Application No. (if known): Attorney Docket No.: 524592007100

Certificate of Express Mailing Under 37 CFR 1.10

I hereby certify that this correspondence is being deposited with the United States Postal Service as Express Mail, Airbill No. EV 272142882 US in an envelope addressed to:

MS Patent Application
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

on November 25, 2003
Date

Signature

Bruce D. Grant
Typed or printed name of person signing Certificate

Note: Each paper must have its own certificate of mailing, or this certificate must identify each submitted paper.

Application Data Sheet

sd-172881

METHODS FOR IDENTIFYING RISK OF BREAST CANCER AND TREATMENTS THEREOF

Related Patent Applications

[0001] This patent application claims the benefit of provisional patent application no. 60/429,136 filed November 25, 2002 and provisional patent application no. 60/490,234 filed July 24, 2003, having attorney docket number 524593004100 and 524593004101, respectively. Each of these provisional patent applications names Richard B. Roth *et al.* as inventors and is hereby incorporated herein by reference in its entirety, including all drawings and cited publications and documents.

Field of the Invention

[0002] The invention relates to genetic methods for identifying risk of breast cancer and treatments that specifically target the disease.

Background

[0003] Breast cancer is the third most common cancer, and the most common cancer in women, as well as a cause of disability, psychological trauma, and economic loss. Breast cancer is the second most common cause of cancer death in women in the United States, in particular for women between the ages of 15 and 54, and the leading cause of cancer-related death (Forbes, Seminars in Oncology, vol.24(1), Suppl 1, 1997: pp.S1-20-S1-35). Indirect effects of the disease also contribute to the mortality from breast cancer including consequences of advanced disease, such as metastases to the bone or brain. Complications arising from bone marrow suppression, radiation fibrosis and neutropenic sepsis, collateral effects from therapeutic interventions, such as surgery, radiation, chemotherapy, or bone marrow transplantation-also contribute to the morbidity and mortality from this disease.

[0004] While the pathogenesis of breast cancer is unclear, transformation of normal breast epithelium to a malignant phenotype may be the result of genetic factors, especially in women under thirty (Miki, *et al.*, Science, 266: 66-71 (1994)). However, it is likely that other, non-genetic factors also have a significant effect on the etiology of the disease. Regardless of its origin, breast cancer morbidity increases significantly if it is not detected early in its progression. Thus, considerable efforts have focused on the elucidation of early cellular events surrounding transformation in breast tissue. Such efforts have led to the identification of several potential breast cancer markers. For example, alleles of the *BRCA1* and *BRCA2* genes have been linked to hereditary and early-onset breast cancer (Wooster, *et al.*, Science, 265: 2088-2090 (1994)). However, *BRCA1* is limited as a cancer marker because *BRCA1*

mutations fail to account for the majority of breast cancers (Ford, *et al.*, British J. Cancer, 72: 805-812 (1995)). Similarly, the *BRCA2* gene, which has been linked to forms of hereditary breast cancer, accounts for only a small portion of total breast cancer cases.

Summary

[0005] It has been discovered that certain polymorphic variations in human genomic DNA are associated with the occurrence of breast cancer. In particular, polymorphic variants in loci containing *DLG1, KIAA0783, DPF3* and *CENPC1* regions in human genomic DNA have been associated with risk of breast cancer.

[0006] Thus, featured herein are methods for identifying a subject at risk of breast cancer and/or a risk of breast cancer in a subject, which comprises detecting the presence or absence of one or more polymorphic variations associated with breast cancer in genomic regions described herein in a human nucleic acid sample. In an embodiment, two or more polymorphic variations are detected in two or more regions selected from the group consisting of *DLG1, KIAA0783, DPF3* and *CENPC1*. In certain embodiments, 3 or fewer, or 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 or 20 or fewer polymorphic variants are detected.

[0007] Also featured are nucleic acids that include one or more polymorphic variations associated with the occurrence of breast cancer, as well as polypeptides encoded by these nucleic acids. Further, provided is a method for identifying a subject at risk of breast cancer and then prescribing to the subject a breast cancer detection procedure, prevention procedure and/or a treatment procedure. In addition, provided are methods for identifying candidate therapeutic molecules for treating breast cancer and related disorders, as well as methods for treating breast cancer in a subject by diagnosing breast cancer in the subject and treating the subject with a suitable treatment, such as administering a therapeutic molecule.

[0008] Also provided are compositions comprising a breast cancer cell and/or *DLG1, KIAA0783, DPF3* or *CENPC1* nucleic acid with a RNAi, siRNA, antisense DNA or RNA, or ribozyme nucleic acid designed from a *DLG1, KIAA0783, DPF3* or *CENPC1* nucleotide sequence. In an embodiment, the nucleic acid is designed from a *DLG1, KIAA0783, DPF3* or *CENPC1* nucleotide sequence that includes one or more breast cancer associated polymorphic variations, and in some instances, specifically interacts with such a nucleotide sequence. Further, provided are arrays of nucleic acids bound to a solid surface, in which one or more nucleic acid molecules of the array have a *DLG1, KIAA0783, DPF3* or *CENPC1* nucleotide sequence, or a fragment or substantially identical nucleic acid thereof, or a complementary nucleic acid of the foregoing. Featured also are compositions comprising a breast cancer cell and/or a *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide, with an antibody that specifically binds to the

polypeptide. In an embodiment, the antibody specifically binds to an epitope in the polypeptide that includes a non-synonymous amino acid modification associated with breast cancer (e.g., results in an amino acid substitution in the encoded polypeptide associated with breast cancer). In certain embodiments, the antibody specifically binds to an epitope that comprises a glutamine at amino acid position 278 in SEQ ID NO: 9 of a *DLG1* polypeptide or a glycine at amino acid position 389 in SEQ ID NO: 12 of a *CENPC1* polypeptide.

Brief Description of the Figures

[0009] Figures 1A-1T show a genomic nucleotide sequence for an *DLG1* region. The genomic nucleotide sequence is set forth in SEQ ID NO: 1. The following nucleotide representations are used throughout: "A" or "a" is adenosine, adenine, or adenylic acid; "C" or "c" is cytidine, cytosine, or cytidylic acid; "G" or "g" is guanosine, guanine, or guanylic acid; "T" or "t" is thymidine, thymine, or thymidylic acid; and "I" or "i" is inosine, hypoxanthine, or inosinic acid. Exons are indicated in italicized lower case type, introns are depicted in normal text lower case type, and polymorphic sites are depicted in bold upper case type. SNPs are designated by the following convention: "R" represents A or G, "M" represents A or C; "W" represents A or T; "Y" represents C or T; "S" represents C or G; "K" represents G or T; "V" represents A, C or G; "H" represents A, C, or T; "D" represents A, G, or T; "B" represents C, G, or T; and "N" represents A, G, C, or T.

[0010] Figures 2A-2Z show a genomic nucleotide sequence of a *KIAA0783* region. The genomic nucleotide sequence is set forth in SEQ ID NO: 2.

[0011] Figures 3A-3X show a genomic nucleotide sequence of a *DPF3* region. The genomic nucleotide sequence is set forth in SEQ ID NO: 3.

[0012] Figures 4A-4Y show a genomic nucleotide sequence of a *CENPC1* region. The genomic nucleotide sequence is set forth in SEQ ID NO: 4.

[0013] Figure 5 shows a coding nucleotide sequence (cDNA) for *DLG1*. The nucleotide sequence is set forth in SEQ ID NO: 5.

[0014] Figure 6 shows a coding nucleotide sequence (cDNA) for *KIAA0783*. The nucleotide sequence is set forth in SEQ ID NO: 6.

[0015] Figure 7 shows a coding nucleotide sequence (cDNA) for *DPF3*. The nucleotide sequence is set forth in SEQ ID NO: 7.

[0016] Figure 8 shows a coding nucleotide sequence (cDNA) for *CENPC1*. The nucleotide sequence is set forth in SEQ ID NO: 8.

[0017] Figure 9 shows an amino acid sequence for a *DLG1* polypeptide, which is set forth in SEQ ID NO: 9.

[0018] Figure 10 shows an amino acid sequence for a *KIAA0783* polypeptide, which is set forth in SEQ ID NO: 10.

[0019] Figure 11 shows an amino acid sequence for a *DPF3* polypeptide, which is set forth in SEQ ID NO: 11.

[0020] Figure 12 shows an amino acid sequence for a *CENPC1* polypeptide, which is set forth in SEQ ID NO: 12.

[0021] Figures 13-16 show proximal SNPs in *DLG1, KIAA0783, DPF3* and *CENPC1* loci in genomic DNA. The position of each SNP on the chromosome is shown on the x-axis and the y-axis provides the negative logarithm of the p-value comparing the estimated allele to that of the control group. Also shown in the figure are exons and introns of the genes in the approximate chromosomal positions. The figure indicates that polymorphic variants associated with breast cancer are in linkage disequilibrium in the following regions: the region spanning positions 7938-59808 in SEQ ID NO: 1; the region spanning positions 10511-98107 in SEQ ID NO: 2; the region spanning positions 160-72752 in SEQ ID NO: 3; and the region spanning positions 196-74909 in SEQ ID NO: 4.

Detailed Description

[0022] It has been discovered that polymorphic variations in the *DLG1, KIAA0783, DPF3* and *CENPC1* regions described herein are associated with an increased risk of breast cancer.

[0023] The gene *DLG1* (discs, large homolog 1 (Drosophila)) is also referenced as synapse-associated protein 97, hdlg, SAP97. *DLG1* has been mapped to chromosomal position 3-q29. In Drosophila more than 50 genes have been identified that lead to loss of cell proliferation control, indicating that they are tumor suppressor genes. Many of these genes have been cloned and sequenced, and most have clear mammalian homologs. The Drosophila 'discs large' tumor suppressor protein, Dlg, is the prototype of a family of proteins termed MAGUKs (membrane-associated guanylate kinase homologs). MAGUKs are localized at the membrane-cytoskeleton interface, usually at cell-cell junctions, where they appear to have both structural and signaling roles. They contain several distinct domains, including a modified guanylate kinase domain, an SH3 motif, and 1 or 3 copies of the DHR (GLGF/PDZ) domain. Recessive lethal mutations in the 'discs large' tumor suppressor gene interfere with the formation of septate junctions (thought to be the arthropod equivalent of tight junctions) between epithelial cells, and they also cause neoplastic overgrowth of imaginal discs, suggesting a role for cell junctions in proliferation control.

[0024] The gene *KIAA0783* also is known as PHF14 and PHD finger protein 14. *KIAA0783* has been mapped to chromosomal position 7p21.3. The protein encoded by this gene is a novel gene with unknown function. Being a zinc finger protein, it likely a transcription factor.

[0025] The gene *DPF3* (D4, zinc and double PHD fingers, family 3) also is known as CERD4, cer-d4, FLJ14079, and 2810403B03Rik. DPF3 is a Rho family guanine-nucleotide exchange factor. *DPF3* has been mapped to chromosomal position 14q24.3-q31.1.

[0026] The gene *CENPC1* (centromere protein C1) also is known as Centromere autoantigen C1. CENPC1 has been mapped to chromosomal position 4q12-q13.3. *CENPC1* is a centromere autoantigen and a component of the inner kinetochore plate. The protein is required for maintaining proper kinetochore size and a timely transition to anaphase. A putative pseudogene exists on chromosome 12.

Breast Cancer and Sample Selection

[0027] Breast cancer is typically described as the uncontrolled growth of malignant breast tissue. Breast cancers arise most commonly in the lining of the milk ducts of the breast (ductal carcinoma), or in the lobules where breast milk is produced (lobular carcinoma). Other forms of breast cancer include Inflammatory Breast Cancer and Recurrent Breast Cancer. Inflammatory breast cancer is a rare, but very serious, aggressive type of breast cancer. The breast may look red and feel warm with ridges, welts, or hives on the breast; or the skin may look wrinkled. It is sometimes misdiagnosed as a simple infection. Recurrent disease means that the cancer has come back after it has been treated. It may come back in the breast, in the soft tissues of the chest (the chest wall), or in another part of the body.

[0028] As used herein, the term "breast cancer" refers to a condition characterized by anomalous rapid proliferation of abnormal cells in one or both breasts of a subject. The abnormal cells often are referred to as "neoplastic cells," which are transformed cells that can form a solid tumor. The term "tumor" refers to an abnormal mass or population of cells (*i.e.* two or more cells) that result from excessive or abnormal cell division, whether malignant or benign, and pre-cancerous and cancerous cells. Malignant tumors are distinguished from benign growths or tumors in that, in addition to uncontrolled cellular proliferation, they can invade surrounding tissues and can metastasize. In breast cancer, neoplastic cells may be identified in one or both breasts only and not in another tissue or organ, in one or both breasts and one or more adjacent tissues or organs (*e.g.* lymph node), or in a breast and one or more non-adjacent tissues or organs to which the breast cancer cells have metastasized.

[0029] The term "invasion" as used herein refers to the spread of cancerous cells to adjacent surrounding tissues. The term "invasion" often is used synonymously with the term "metastasis," which as used herein refers to a process in which cancer cells travel from one organ or tissue to another non-adjacent organ or tissue. Cancer cells in the breast(s) can spread to tissues and organs of a subject, and conversely, cancer cells from other organs or tissue can invade or metastasize to a breast. Cancerous cells from the breast(s) may invade or metastasize to any other organ or tissue of the body. Breast cancer

cells often invade lymph node cells and/or metastasize to the liver, brain and/or bone and spread cancer in these tissues and organs. Breast cancers can spread to other organs and tissues and cause lung cancer, prostate cancer, colon cancer, ovarian cancer, cervical cancer, gastrointestinal cancer, pancreatic cancer, glioblastoma, bladder cancer, hepatoma, colorectal cancer, uterine cervical cancer, endometrial carcinoma, salivary gland carcinoma, kidney cancer, vulval cancer, thyroid cancer, hepatic carcinoma, skin cancer, melanoma, ovarian cancer, neuroblastoma, myeloma, various types of head and neck cancer, acute lymphoblastic leukemia, acute myeloid leukemia, Ewing sarcoma and peripheral neuroepithelioma, and other carcinomas, lymphomas, blastomas, sarcomas, and leukemias.

[0030] Breast cancers arise most commonly in the lining of the milk ducts of the breast (ductal carcinoma), or in the lobules where breast milk is produced (lobular carcinoma). Other forms of breast cancer include Inflammatory Breast Cancer and Recurrent Breast Cancer. Inflammatory Breast Cancer is a rare, but very serious, aggressive type of breast cancer. The breast may look red and feel warm with ridges, welts, or hives on the breast; or the skin may look wrinkled. It is sometimes misdiagnosed as a simple infection. Recurrent disease means that the cancer has come back after it has been treated. It may come back in the breast, in the soft tissues of the chest (the chest wall), or in another part of the body. As used herein, the term "breast cancer" may include both Inflammatory Breast Cancer and Recurrent Breast Cancer.

[0031] In an effort to detect breast cancer as early as possible, regular physical exams and screening mammograms often are prescribed and conducted. A diagnostic mammogram often is performed to evaluate a breast complaint or abnormality detected by physical exam or routine screening mammography. If an abnormality seen with diagnostic mammography is suspicious, additional breast imaging (with exams such as ultrasound) or a biopsy may be ordered. A biopsy followed by pathological (microscopic) analysis is a definitive way to determine whether a subject has breast cancer. Excised breast cancer samples often are subjected to the following analyses: diagnosis of the breast tumor and confirmation of its malignancy; maximum tumor thickness; assessment of completeness of excision of invasive and *in situ* components and microscopic measurements of the shortest extent of clearance; level of invasion; presence and extent of regression; presence and extent of ulceration; histological type and special variants; pre-existing lesion; mitotic rate; vascular invasion; neurotropism; cell type; tumor lymphocyte infiltration; and growth phase.

[0032] The stage of a breast cancer can be classified as a range of stages from Stage 0 to Stage IV based on its size and the extent to which it has spread. The following table summarizes the stages:

Table A

Stage	Tumor Size	Lymph Node Involvement	Metastasis (Spread)
I	Less than 2 cm	No	No
II	Between 2-5 cm	No or in same side of breast	No
III	More than 5 cm	Yes, on same side of breast	No
IV	Not applicable	Not applicable	Yes

[0033] Stage 0 cancer is a contained cancer that has not spread beyond the breast ductal system. Fifteen to twenty percent of breast cancers detected by clinical examinations or testing are in Stage 0 (the earliest form of breast cancer). Two types of Stage 0 cancer are lobular carcinoma in situ (LCIS) and ductal carcinoma in situ (DCIS). LCIS indicates high risk for breast cancer. Many physicians do not classify LCIS as a malignancy and often encounter LCIS by chance on breast biopsy while investigating another area of concern. While the microscopic features of LCIS are abnormal and are similar to malignancy, LCIS does not behave as a cancer (and therefore is not treated as a cancer). LCIS is merely a marker for a significantly increased risk of cancer anywhere in the breast. However, bilateral simple mastectomy may be occasionally performed if LCIS patients have a strong family history of breast cancer. In DCIS the cancer cells are confined to milk ducts in the breast and have not spread into the fatty breast tissue or to any other part of the body (such as the lymph nodes). DCIS may be detected on mammogram as tiny specks of calcium (known as microcalcifications) 80% of the time. Less commonly DCIS can present itself as a mass with calcifications (15% of the time); and even less likely as a mass without calcifications (<5% of the time). A breast biopsy is used to confirm DCIS. A standard DCIS treatment is breast-conserving therapy (BCT), which is lumpectomy followed by radiation treatment or mastectomy. To date, DCIS patients have chosen equally among lumpectomy and mastectomy as their treatment option, though specific cases may sometimes favor lumpectomy over mastectomy or vice versa.

[0034] In Stage I, the primary (original) cancer is 2 cm or less in diameter and has not spread to the lymph nodes. In Stage IIA, the primary tumor is between 2 and 5 cm in diameter and has not spread to the lymph nodes. In Stage IIB, the primary tumor is between 2 and 5 cm in diameter and has spread to the axillary (underarm) lymph nodes; or the primary tumor is over 5 cm and has not spread to the lymph nodes. In Stage IIIA, the primary breast cancer of any kind that has spread to the axillary (underarm) lymph nodes and to axillary tissues. In Stage IIIB, the primary breast cancer is any size, has attached

itself to the chest wall, and has spread to the pectoral (chest) lymph nodes. In Stage IV, the primary cancer has spread out of the breast to other parts of the body (such as bone, lung, liver, brain). The treatment of Stage IV breast cancer focuses on extending survival time and relieving symptoms.

[0035] Based in part upon selection criteria set forth above, individuals having breast cancer can be selected for genetic studies. Also, individuals having no history of cancer or breast cancer often are selected for genetic studies. Other selection criteria can include: a tissue or fluid sample is derived from an individual characterized as Caucasian; the sample was derived from an individual of German paternal and maternal descent; the database included relevant phenotype information for the individual; case samples were derived from individuals diagnosed with breast cancer; control samples were derived from individuals free of cancer and no family history of breast cancer; and sufficient genomic DNA was extracted from each blood sample for all allelotyping and genotyping reactions performed during the study. Phenotype information included pre- or post-menopausal, familial predisposition, country or origin of mother and father, diagnosis with breast cancer (date of primary diagnosis, age of individual as of primary diagnosis, grade or stage of development, occurrence of metastases, *e.g.*, lymph node metastases, organ metastases), condition of body tissue (skin tissue, breast tissue, ovary tissue, peritoneum tissue and myometrium), method of treatment (surgery, chemotherapy, hormone therapy, radiation therapy).

[0036] Provided herein is a set of blood samples and a set of corresponding nucleic acid samples isolated from the blood samples, where the blood samples are donated from individuals diagnosed with breast cancer. The sample set often includes blood samples or nucleic acid samples from 100 or more, 150 or more, or 200 or more individuals having breast cancer, and sometimes from 250 or more, 300 or more, 400 or more, or 500 or more individuals. The individuals can have parents from any place of origin, and in an embodiment, the set of samples are extracted from individuals of German paternal and German maternal ancestry. The samples in each set may be selected based upon five or more criteria and/or phenotypes set forth above.

Polymorphic Variants Associated with Breast Cancer

[0037] A genetic analysis provided herein linked breast cancer with polymorphic variants in the *DLG1, KIAA0783, DPF3* and *CENPC1* regions of the human genome disclosed herein. As used herein, the term "polymorphic site" refers to a region in a nucleic acid at which two or more alternative nucleotide sequences are observed in a significant number of nucleic acid samples from a population of individuals. A polymorphic site may be a nucleotide sequence of two or more nucleotides, an inserted nucleotide or nucleotide sequence, a deleted nucleotide or nucleotide sequence, or a microsatellite, for example. A polymorphic site that is two or more nucleotides in length may be 3, 4, 5, 6, 7, 8, 9, 10, 11,

12, 13, 14, 15 or more, 20 or more, 30 or more, 50 or more, 75 or more, 100 or more, 500 or more, or about 1000 nucleotides in length, where all or some of the nucleotide sequences differ within the region. A polymorphic site is often one nucleotide in length, which is referred to herein as a "single nucleotide polymorphism" or a "SNP."

[0038] Where there are two, three, or four alternative nucleotide sequences at a polymorphic site, each nucleotide sequence is referred to as a "polymorphic variant" or "nucleic acid variant." Where two polymorphic variants exist, for example, the polymorphic variant represented in a minority of samples from a population is sometimes referred to as a "minor allele" and the polymorphic variant that is more prevalently represented is sometimes referred to as a "major allele." Many organisms possess a copy of each chromosome (*e.g.*, humans), and those individuals who possess two major alleles or two minor alleles are often referred to as being "homozygous" with respect to the polymorphism, and those individuals who possess one major allele and one minor allele are normally referred to as being "heterozygous" with respect to the polymorphism. Individuals who are homozygous with respect to one allele are sometimes predisposed to a different phenotype as compared to individuals who are heterozygous or homozygous with respect to another allele.

[0039] Furthermore, a genotype or polymorphic variant may be expressed in terms of a "haplotype," which as used herein refers to two or more polymorphic variants occurring within genomic DNA in a group of individuals within a population. For example, two SNPs may exist within a gene where each SNP position includes a cytosine variation and an adenine variation. Certain individuals in a population may carry one allele (heterozygous) or two alleles (homozygous) having the gene with a cytosine at each SNP position. As the two cytosines corresponding to each SNP in the gene travel together on one or both alleles in these individuals, the individuals can be characterized as having a cytosine/cytosine haplotype with respect to the two SNPs in the gene.

[0040] As used herein, the term "phenotype" refers to a trait which can be compared between individuals, such as presence or absence of a condition, a visually observable difference in appearance between individuals, metabolic variations, physiological variations, variations in the function of biological molecules, and the like. An example of a phenotype is occurrence of breast cancer.

[0041] Researchers sometimes report a polymorphic variant in a database without determining whether the variant is represented in a significant fraction of a population. Because a subset of these reported polymorphic variants are not represented in a statistically significant portion of the population, some of them are sequencing errors and/or not biologically relevant. Thus, it is often not known whether a reported polymorphic variant is statistically significant or biologically relevant until the presence of the variant is detected in a population of individuals and the frequency of the variant is determined. Methods for detecting a polymorphic variant in a population are described herein, specifically in Example 2. A

polymorphic variant is statistically significant and often biologically relevant if it is represented in 5% or more of a population, sometimes 10% or more, 15% or more, or 20% or more of a population, and often 25% or more, 30% or more, 35% or more, 40% or more, 45% or more, or 50% or more of a population.

[0042] A polymorphic variant may be detected on either or both strands of a double-stranded nucleic acid. For example, a thymine at a particular position in SEQ ID NO: 1 can be reported as an adenine from the complementary strand. Also, a polymorphic variant may be located within an intron or exon of a gene or within a portion of a regulatory region such as a promoter, a 5′ untranslated region (UTR), a 3′ UTR, and in DNA (*e.g.*, genomic DNA (gDNA) and complementary DNA (cDNA)), RNA (*e.g.*, mRNA, tRNA, and rRNA), or a polypeptide. Polymorphic variations may or may not result in detectable differences in gene expression, polypeptide structure, or polypeptide function.

[0043] In the genetic analysis that associated breast cancer with the polymorphic variants described hereafter, samples from individuals having breast cancer and individuals not having cancer were allelotyped and genotyped. The term "genotyped" as used herein refers to a process for determining a genotype of one or more individuals, where a "genotype" is a representation of one or more polymorphic variants in a population. Genotypes may be expressed in terms of a "haplotype," which as used herein refers to two or more polymorphic variants occurring within genomic DNA in a group of individuals within a population. For example, two SNPs may exist within a gene where each SNP position includes a cytosine variation and an adenine variation. Certain individuals in a population may carry one allele (heterozygous) or two alleles (homozygous) having the gene with a cytosine at each SNP position. As the two cytosines corresponding to each SNP in the gene travel together on one or both alleles in these individuals, the individuals can be characterized as having a cytosine/cytosine haplotype with respect to the two SNPs in the gene.

[0044] It was determined that polymorphic variations associated with an increased risk of breast cancer existed in *DLG1, KIAA0783, DPF3* or *CENPC1* nucleotide sequences. Polymorphic variants in and around the *DLG1, KIAA0783, DPF3* and *CENPC1* loci were tested for association with breast cancer. In the *DLG1* locus, these included polymorphic variants at positions in SEQ ID NO: 1 selected from the group consisting of 133, 7938, 8873, 13221, 17288, 25732, 26923, 39977, 41284, 41410, 41477, 41514, 42606, 42742, 59515, 59808, 60265, 67152, 68332, 71128 and 76427. Polymorphic variants in a region spanning positions 7938-59808 in SEQ ID NO: 1 in particular were associated with an increased risk of breast cancer, including polymorphic variants at positions 7938, 26923, 39977 and 59808 in SEQ ID NO: 1. At these positions in SEQ ID NO: 1, a thymine at position 7938, a cytosine at position 26923, a thymine at position 39977 and a thymine at position 59808 in particular were associated with risk of breast cancer. Also, a glutamine at position 278 in SEQ ID NO: 9 in a *DLG1* polypeptide in particular was associated with an increased risk of breast cancer.

[0045] In the *KIAA0783* locus, these included polymorphic variants at positions in SEQ ID NO: 2 selected from the group consisting of 201, 6395, 8558, 9429, 9809, 10072, 10511, 11556, 16857, 16951, 17027, 17177, 17615, 17950, 18329, 18384, 18561, 18579, 18871, 27152, 27306, 28091, 28661, 29011, 29962, 29969, 30085, 31656, 31685, 31749, 45389, 45459, 46647, 49860, 53061, 57308, 61563, 61660, 62212, 67090, 67198, 70071, 70191, 74006, 75600, 85761, 90798, 90883, 91259, 95416, 95446, 96368, 97050, 97362, 97630, 97989 and 98107. Polymorphic variants in a region spanning positions 10511-98107 in SEQ ID NO: 2 in particular were associated with an increased risk of breast cancer, including polymorphic variants at positions 10511, 11556, 17177, 18384, 28661, 31656, 31685, 31749, 45389, 45459, 46647, 49860, 53061, 57308, 61563, 61660, 67090, 67198, 70071, 74006, 75600, 85761, 90798, 90883, 91259, 95416, 95446, 96368, 97362, 97630, 97989 and 98107 in SEQ ID NO: 2. At these positions in SEQ ID NO: 2, a thymine at position 10511, a cytosine at position 11556, a thymine at position 17177, a thymine at position 18384, an adenine at position 28661, an adenine at position 31656, an adenine at position 31685, a guanine at position 31749, a thymine at position 45389, a guanine at position 45459, an adenine at position 46647, a thymine at position 49860, a thymine at position 53061, an adenine at position 57308, a guanine at position 61563, a guanine at position 61660, a guanine at position 67090, a cytosine at position 67198, an adenine at position 70071, a cytosine at position 74006, an adenine at position 75600, a guanine at position 85761, a thymine at poisition 90798, a cytosine at position 90883, an adenine at position 91259, a cytosine at position 95416, a thymine at position 95446, a thymine at position 96368, a thymine at position 97362, an adenine at position 97630, a cytosine at position 97989 and a thymine at position 98107 in particular were associated with increased risk of breast cancer.

[0046] In the *DPF3* locus, these included polymorphic variants at positions in SEQ ID NO: 3 selected from the group consisting of 160, 6053, 9719, 10481, 10676, 17179, 18561, 18658, 18694, 18858, 24582, 24683, 24767, 27402, 28150, 28494, 32003, 35588, 35619, 35856, 36254, 37314, 40033, 40095, 42593, 42799, 43090, 46683, 49774, 51796, 52079, 53857, 53971, 55899, 60682, 61291, 72720, 72752, 85507 and 89751. Polymorphic variants in a region spanning positions 160-72752 in SEQ ID NO: 3 in particular were associated with an increased risk of breast cancer, including polymorphic variants at positions 160, 6053, 18658, 18694, 18858, 24683, 27402, 28494, 32003, 35588, 35856, 40095, 46683, 52079, 53857, 72720 and 72752 in SEQ ID NO: 3. At these positions in SEQ ID NO: 3, an adenine at position 160, a guanine at position 6053, a guanine at position 18658, a guanine at position 18694, a thymine at position 18858, a guanine at position 24683, a guanine at position 27402, a thymine at position 28494, an adenine at position 32003, a cytosine at position 35588, an adenine at position 35856, a guanine at position 40095, an adenine at position 46683, an adenine at position 52079, a

cytosine at position 53857, an adenine at position 72720 and a cytosine at position 72752 in particular were associated with an increased risk of breast cancer.

[0047] In the *CENPC1* locus, these included polymorphic variants at positions in SEQ ID NO: 4 selected from the group consisting of 196, 13311, 14486, 14691, 15551, 17702, 17872, 19588, 19910, 20006, 20575, 21092, 22830, 23455, 23716, 23890, 24001, 24995, 27282, 27779, 29099, 31185, 33994, 34942, 35137, 36538, 37139, 37358, 38828, 39469, 40233, 40472, 41679, 41682, 42831, 42976, 44128, 44195, 46769, 47363, 48843, 52574, 52602, 53212, 53781, 54710, 55808, 57987, 58556, 59148, 59286, 60217, 60412, 60753, 60791, 61524, 62543, 62825, 62826, 62857, 63400, 63960, 64307, 64539, 65728, 66000, 66521, 68185, 69643, 74909, 82973, 83039, 85713, 86873, 90293, 91810, 92609, 92884 and 42831. Polymorphic variants in a region spanning positions 196-74909 in SEQ ID NO: 4 in particular were associated with an increased risk of breast cancer, including polymorphic variants at positions 196, 13311, 14486, 19910, 20575, 23716, 23890, 24995, 29099, 33994, 34942, 37139, 40233, 40472,42831, 42976, 44195, 48843, 58556, 59286, 60217, 62826, 62857, 63400, 63960 and 74909 in SEQ ID NO: 4. At these positions in SEQ ID NO: 4, an adenine at position 196, a guanine at position 13311, a thymine at position 14486, a thymine at position 19910, an adenine at position 20575, a guanine at position 23716, a guanine at position 23890, an adenine at position 24995, a cytosine at position 29099, a thymine at position 33994, a thymine at position 34942, a thymine at position 37139, a thymine at position 40233, an adenine at position 40472, a guanine at position 42831, a guanine at position 42976, a thymine at position 44195, a thymine at position 48843, an adenine at position 58556, a guanine at position 59286, an adenine at position 60217, a cytosine at position 62826, a thymine at position 62857, a thymine at position 63400, an adenine at position 63960 and a cytosine at position 74909 in particular were associated with an increased risk of breast cancer. Also, a glycine at position 389 in SEQ ID NO: 12 in a *CENPC1* polypeptide in particular was associated with an increased risk of breast cancer.

Additional Polymorphic Variants Associated with Breast Cancer

[0048] Also provided is a method for identifying polymorphic variants proximal to an incident, founder polymorphic variant associated with breast cancer. Thus, featured herein are methods for identifying a polymorphic variation associated with breast cancer that is proximal to an incident polymorphic variation associated with breast cancer, which comprises identifying a polymorphic variant proximal to the incident polymorphic variant associated with breast cancer, where the incident polymorphic variant is in a nucleotide sequence set forth in SEQ ID NO: 1-4. The nucleotide sequence often comprises a polynucleotide sequence selected from the group consisting of (a) a nucleotide sequence set forth in SEQ ID NO: 1-4; (b) a nucleotide sequence which encodes a polypeptide having an amino acid sequence encoded by a nucleotide sequence in SEQ ID NO: 1-4; (c) a nucleotide sequence

which encodes a polypeptide that is 90% or more identical to an amino acid sequence encoded by a nucleotide sequence in SEQ ID NO: 1-4 or a nucleotide sequence about 90% or more identical to the nucleotide sequence set forth in SEQ ID NO: 1-4; and (d) a fragment of a nucleotide sequence of (a), (b), or (c), often a fragment that includes a polymorphic site associated with breast cancer. The presence or absence of an association of the proximal polymorphic variant with breast cancer then is determined using a known association method, such as a method described in the Examples hereafter. In an embodiment, the incident polymorphic variant is described in SEQ ID NO: 1-4. In another embodiment, the proximal polymorphic variant identified sometimes is a publicly disclosed polymorphic variant, which for example, sometimes is published in a publicly available database. In other embodiments, the polymorphic variant identified is not publicly disclosed and is discovered using a known method, including, but not limited to, sequencing a region surrounding the incident polymorphic variant in a group of nucleic acid samples. Thus, multiple polymorphic variants proximal to an incident polymorphic variant are associated with breast cancer using this method.

[0049] The proximal polymorphic variant often is identified in a region surrounding the incident polymorphic variant. In certain embodiments, this surrounding region is about 50 kb flanking the first polymorphic variant (*e.g.* about 50 kb 5' of the first polymorphic variant and about 50 kb 3' of the first polymorphic variant), and the region sometimes is composed of shorter flanking sequences, such as flanking sequences of about 40 kb, about 30 kb, about 25 kb, about 20 kb, about 15 kb, about 10 kb, about 7 kb, about 5 kb, or about 2 kb 5' and 3' of the incident polymorphic variant. In other embodiments, the region is composed of longer flanking sequences, such as flanking sequences of about 55 kb, about 60 kb, about 65 kb, about 70 kb, about 75 kb, about 80 kb, about 85 kb, about 90 kb, about 95 kb, or about 100 kb 5' and 3' of the incident polymorphic variant.

[0050] In certain embodiments, polymorphic variants associated with breast cancer are identified iteratively. For example, a first proximal polymorphic variant is associated with breast cancer using the methods described above and then another polymorphic variant proximal to the first proximal polymorphic variant is identified (*e.g.*, publicly disclosed or discovered) and the presence or absence of an association of one or more other polymorphic variants proximal to the first proximal polymorphic variant with breast cancer is determined.

[0051] The methods described herein are useful for identifying or discovering additional polymorphic variants that may be used to further characterize a gene, region or loci associated with a condition, a disease (*e.g.*, breast cancer), or a disorder. For example, allelotyping or genotyping data from the additional polymorphic variants may be used to identify a functional mutation or a region of linkage disequilibrium.

[0052] In certain embodiments, polymorphic variants identified or discovered within a region comprising the first polymorphic variant associated with breast cancer are genotyped using the genetic methods and sample selection techniques described herein, and it can be determined whether those polymorphic variants are in linkage disequilibrium with the first polymorphic variant. The size of the region in linkage disequilibrium with the first polymorphic variant also can be assessed using these genotyping methods. Thus, provided herein are methods for determining whether a polymorphic variant is in linkage disequilibrium with a first polymorphic variant associated with breast cancer, and such information can be used in prognosis methods described herein.

Isolated *DLG1, KIAA0783, DPF3* or *CENPC1* Nucleic Acids

[0053] Featured herein are isolated *DLG1, KIAA0783, DPF3* or *CENPC1* nucleic acids, which include the nucleic acid having the nucleotide sequence of SEQ ID NO: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 or 11, nucleic acid variants, and substantially identical nucleic acids of the foregoing. Nucleotide sequences of the *DLG1, KIAA0783, DPF3* or *CENPC1* nucleic acids sometimes are referred to herein as "*DLG1, KIAA0783, DPF3* or *CENPC1* nucleotide sequences." A "*DLG1, KIAA0783, DPF3* or *CENPC1* nucleic acid variant" refers to one allele that may have one or more different polymorphic variations as compared to another allele in another subject or the same subject. A polymorphic variation in the *DLG1, KIAA0783, DPF3* or *CENPC1* nucleic acid variant may be represented on one or both strands in a double-stranded nucleic acid or on one chromosomal complement (heterozygous) or both chromosomal complements (homozygous).

[0054] As used herein, the term "nucleic acid" includes DNA molecules (e.g., a complementary DNA (cDNA) and genomic DNA (gDNA)) and RNA molecules (e.g., mRNA, rRNA, and tRNA) and analogs of DNA or RNA, for example, by use of nucleotide analogs. The nucleic acid molecule can be single-stranded and it is often double-stranded. The term "isolated or purified nucleic acid" refers to nucleic acids that are separated from other nucleic acids present in the natural source of the nucleic acid. For example, with regard to genomic DNA, the term "isolated" includes nucleic acids which are separated from the chromosome with which the genomic DNA is naturally associated. An "isolated" nucleic acid is often free of sequences which naturally flank the nucleic acid (i.e., sequences located at the 5' and/or 3' ends of the nucleic acid) in the genomic DNA of the organism from which the nucleic acid is derived. For example, in various embodiments, the isolated nucleic acid molecule can contain less than about 5 kb, 4 kb, 3 kb, 2 kb, 1 kb, 0.5 kb or 0.1 kb of 5' and/or 3' nucleotide sequences which flank the nucleic acid molecule in genomic DNA of the cell from which the nucleic acid is derived. Moreover, an "isolated" nucleic acid molecule, such as a cDNA molecule, can be substantially free of other cellular material, or culture medium when produced by recombinant techniques, or substantially free of chemical

precursors or other chemicals when chemically synthesized. As used herein, the term "*DLG1, KIAA0783, DPF3* or *CENPC1* gene" refers to a nucleotide sequence that encodes a *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide.

[0055] Also included herein are nucleic acid fragments. These fragments typically are a nucleotide sequence identical to a nucleotide sequence in SEQ ID NO: 1-8, a nucleotide sequence substantially identical to a nucleotide sequence in SEQ ID NO: 1-8, or a nucleotide sequence that is complementary to the foregoing. The nucleic acid fragment may be identical, substantially identical or homologous to a nucleotide sequence in an exon or an intron in SEQ ID NO: 1-4, and may encode a domain or part of a domain or motif of a *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide, sometimes the domains set forth in Figures 13-18. Sometimes, the fragment comprises the polymorphic variation described herein as being associated with breast cancer. The nucleic acid fragment sometimes is 50, 100, or 200 or fewer base pairs in length, and is sometimes about 300, 400, 500, 600, 700, 800, 900, 1000, 1100, 1200, 1300, 1400, 1500, 1600, 1700, 1800, 1900, 2000, 2100, 2200, 2300, 2400, 2500, 2600, 2700, 2800, 2900, 3000, 3100, 3200, 3300, 3400, 3500, 3600, 3800, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 15000, 20000, 30000, 40000, 50000, 60000, 70000, 80000, 90000, 100000, 110000, 120000, 130000, 140000, 150000 or 160000 base pairs in length. A nucleic acid fragment complementary to a nucleotide sequence identical or substantially identical to the nucleotide sequence of SEQ ID NO: 1-8 and hybridizes to such a nucleotide sequence under stringent conditions often is referred to as a "probe." Nucleic acid fragments often include one or more polymorphic sites, or sometimes have an end that is adjacent to a polymorphic site as described hereafter.

[0056] An example of a nucleic acid fragment is an oligonucleotide. As used herein, the term "oligonucleotide" refers to a nucleic acid comprising about 8 to about 50 covalently linked nucleotides, often comprising from about 8 to about 35 nucleotides, and more often from about 10 to about 25 nucleotides. The backbone and nucleotides within an oligonucleotide may be the same as those of naturally occurring nucleic acids, or analogs or derivatives of naturally occurring nucleic acids, provided that oligonucleotides having such analogs or derivatives retain the ability to hybridize specifically to a nucleic acid comprising a targeted polymorphism. Oligonucleotides described herein may be used as hybridization probes or as components of prognostic or diagnostic assays, for example, as described herein.

[0057] Oligonucleotides are typically synthesized using standard methods and equipment, such as the ABI 3900 High Throughput DNA Synthesizer and the EXPEDITE™ 8909 Nucleic Acid Synthesizer, both of which are available from Applied Biosystems (Foster City, CA). Analogs and derivatives are exemplified in U.S. Pat. Nos. 4,469,863; 5,536,821; 5,541,306; 5,637,683; 5,637,684; 5,700,922; 5,717,083; 5,719,262; 5,739,308; 5,773,601; 5,886,165; 5,929,226; 5,977,296; 6,140,482; WO 00/56746;

WO 01/14398, and related publications. Methods for synthesizing oligonucleotides comprising such analogs or derivatives are disclosed, for example, in the patent publications cited above and in U.S. Pat. Nos. 5,614,622; 5,739,314; 5,955,599; 5,962,674; 6,117,992; in WO 00/75372; and in related publications.

[0058] Oligonucleotides also may be linked to a second moiety. The second moiety may be an additional nucleotide sequence such as a tail sequence (e.g., a polyadenosine tail), an adapter sequence (e.g., phage M13 universal tail sequence), and others. Alternatively, the second moiety may be a non-nucleotide moiety such as a moiety which facilitates linkage to a solid support or a label to facilitate detection of the oligonucleotide. Such labels include, without limitation, a radioactive label, a fluorescent label, a chemiluminescent label, a paramagnetic label, and the like. The second moiety may be attached to any position of the oligonucleotide, provided the oligonucleotide can hybridize to the nucleic acid comprising the polymorphism.

Uses for Nucleic Acid Sequences

[0059] Nucleic acid coding sequences depicted in SEQ ID NO: 1-8 may be used for diagnostic purposes for detection and control of polypeptide expression. Also, included herein are oligonucleotide sequences such as antisense RNA, small-interfering RNA (siRNA) and DNA molecules and ribozymes that function to inhibit translation of a polypeptide. Antisense techniques and RNA interference techniques are known in the art and are described herein.

[0060] Ribozymes are enzymatic RNA molecules capable of catalyzing the specific cleavage of RNA. The mechanism of ribozyme action involves sequence specific hybridization of the ribozyme molecule to complementary target RNA, followed by a endonucleolytic cleavage. Ribozymes may be engineered hammerhead motif ribozyme molecules that specifically and efficiently catalyze endonucleolytic cleavage of RNA sequences corresponding to or complementary to the nucleotide sequences set forth in SEQ ID NO: 1-8. Specific ribozyme cleavage sites within any potential RNA target are initially identified by scanning the target molecule for ribozyme cleavage sites which include the following sequences, GUA, GUU and GUC. Once identified, short RNA sequences of between fifteen (15) and twenty (20) ribonucleotides corresponding to the region of the target gene containing the cleavage site may be evaluated for predicted structural features such as secondary structure that may render the oligonucleotide sequence unsuitable. The suitability of candidate targets may also be evaluated by testing their accessibility to hybridization with complementary oligonucleotides, using ribonuclease protection assays.

[0061] Antisense RNA and DNA molecules, siRNA and ribozymes may be prepared by any method known in the art for the synthesis of RNA molecules. These include techniques for chemically

synthesizing oligodeoxyribonucleotides well known in the art such as solid phase phosphoramidite chemical synthesis. Alternatively, RNA molecules may be generated by *in vitro* and *in vivo* transcription of DNA sequences encoding the antisense RNA molecule. Such DNA sequences may be incorporated into a wide variety of vectors which incorporate suitable RNA polymerase promoters such as the T7 or SP6 polymerase promoters. Alternatively, antisense cDNA constructs that synthesize antisense RNA constitutively or inducibly, depending on the promoter used, can be introduced stably into cell lines.

[0062] DNA encoding a polypeptide also may have a number of uses for the diagnosis of diseases, including breast cancer, resulting from aberrant expression of a target gene described herein. For example, the nucleic acid sequence may be used in hybridization assays of biopsies or autopsies to diagnose abnormalities of expression or function (*e.g.*, Southern or Northern blot analysis, in situ hybridization assays).

[0063] In addition, the expression of a polypeptide during embryonic development may also be determined using nucleic acid encoding the polypeptide. As addressed, *infra*, production of functionally impaired polypeptide can be the cause of various disease states, such as breast cancer. *In situ* hybridizations using polynucleotide probes may be employed to predict problems related to breast cancer. Further, as indicated, *infra*, administration of human active polypeptide, recombinantly produced as described herein, may be used to treat disease states related to functionally impaired polypeptide. Alternatively, gene therapy approaches may be employed to remedy deficiencies of functional polypeptide or to replace or compete with dysfunctional polypeptide.

Expression Vectors, Host Cells, and Genetically Engineered Cells

[0064] Provided herein are nucleic acid vectors, often expression vectors, which contain a *DLG1, KIAA0783, DPF3* or *CENPC1* nucleic acid. As used herein, the term "vector" refers to a nucleic acid molecule capable of transporting another nucleic acid to which it has been linked and can include a plasmid, cosmid, or viral vector. The vector can be capable of autonomous replication or it can integrate into a host DNA. Viral vectors may include replication defective retroviruses, adenoviruses and adeno-associated viruses for example.

[0065] A vector can include a *DLG1, KIAA0783, DPF3* or *CENPC1* nucleic acid in a form suitable for expression of the nucleic acid in a host cell. The recombinant expression vector typically includes one or more regulatory sequences operatively linked to the nucleic acid sequence to be expressed. The term "regulatory sequence" includes promoters, enhancers and other expression control elements (e.g., polyadenylation signals). Regulatory sequences include those that direct constitutive expression of a nucleotide sequence, as well as tissue-specific regulatory and/or inducible sequences. The design of the expression vector can depend on such factors as the choice of the host cell to be transformed, the level of

expression of polypeptide desired, and the like. Expression vectors can be introduced into host cells to produce *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptides, including fusion polypeptides, encoded by *DLG1, KIAA0783, DPF3* or *CENPC1* nucleic acids.

[0066] Recombinant expression vectors can be designed for expression of *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptides in prokaryotic or eukaryotic cells. For example, *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptides can be expressed in E. coli, insect cells (e.g., using baculovirus expression vectors), yeast cells, or mammalian cells. Suitable host cells are discussed further in Goeddel, Gene Expression Technology: Methods in Enzymology 185, Academic Press, San Diego, CA (1990). Alternatively, the recombinant expression vector can be transcribed and translated in vitro, for example using T7 promoter regulatory sequences and T7 polymerase.

[0067] Expression of polypeptides in prokaryotes is most often carried out in E. coli with vectors containing constitutive or inducible promoters directing the expression of either fusion or non-fusion polypeptides. Fusion vectors add a number of amino acids to a polypeptide encoded therein, usually to the amino terminus of the recombinant polypeptide. Such fusion vectors typically serve three purposes: 1) to increase expression of recombinant polypeptide; 2) to increase the solubility of the recombinant polypeptide; and 3) to aid in the purification of the recombinant polypeptide by acting as a ligand in affinity purification. Often, a proteolytic cleavage site is introduced at the junction of the fusion moiety and the recombinant polypeptide to enable separation of the recombinant polypeptide from the fusion moiety subsequent to purification of the fusion polypeptide. Such enzymes, and their cognate recognition sequences, include Factor Xa, thrombin and enterokinase. Typical fusion expression vectors include pGEX (Pharmacia Biotech Inc; Smith & Johnson, Gene 67: 31-40 (1988)), pMAL (New England Biolabs, Beverly, MA) and pRIT5 (Pharmacia, Piscataway, NJ) which fuse glutathione S-transferase (GST), maltose E binding polypeptide, or polypeptide A, respectively, to the target recombinant polypeptide.

[0068] Purified fusion polypeptides can be used in screening assays and to generate antibodies specific for *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptides. In a therapeutic embodiment, fusion polypeptide expressed in a retroviral expression vector is used to infect bone marrow cells that are subsequently transplanted into irradiated recipients. The pathology of the subject recipient is then examined after sufficient time has passed (e.g., six (6) weeks).

[0069] Expressing the polypeptide in host bacteria with an impaired capacity to proteolytically cleave the recombinant polypeptide is often used to maximize recombinant polypeptide expression (Gottesman, S., Gene Expression Technology: Methods in Enzymology, Academic Press, San Diego, California 185: 119-128 (1990)). Another strategy is to alter the nucleotide sequence of the nucleic acid to be inserted into an expression vector so that the individual codons for each amino acid are those

preferentially utilized in E. coli (Wada et al., Nucleic Acids Res. 20: 2111-2118 (1992)). Such alteration of nucleotide sequences can be carried out by standard DNA synthesis techniques.

[0070] When used in mammalian cells, the expression vector's control functions are often provided by viral regulatory elements. For example, commonly used promoters are derived from polyoma, Adenovirus 2, cytomegalovirus and Simian Virus 40. Recombinant mammalian expression vectors are often capable of directing expression of the nucleic acid in a particular cell type (e.g., tissue-specific regulatory elements are used to express the nucleic acid). Non-limiting examples of suitable tissue-specific promoters include an albumin promoter (liver-specific; Pinkert et al., Genes Dev. 1: 268-277 (1987)), lymphoid-specific promoters (Calame & Eaton, Adv. Immunol. 43: 235-275 (1988)), promoters of T cell receptors (Winoto & Baltimore, EMBO J. 8: 729-733 (1989)) promoters of immunoglobulins (Banerji et al., Cell 33: 729-740 (1983); Queen & Baltimore, Cell 33: 741-748 (1983)), neuron-specific promoters (e.g., the neurofilament promoter; Byrne & Ruddle, Proc. Natl. Acad. Sci. USA 86: 5473-5477 (1989)), pancreas-specific promoters (Edlund et al., Science 230: 912-916 (1985)), and mammary gland-specific promoters (e.g., milk whey promoter; U.S. Patent No. 4,873,316 and European Application Publication No. 264,166). Developmentally-regulated promoters are sometimes utilized, for example, the murine hox promoters (Kessel & Gruss, Science 249: 374-379 (1990)) and the α-fetopolypeptide promoter (Campes & Tilghman, Genes Dev. 3: 537-546 (1989)).

[0071] A *DLG1, KIAA0783, DPF3* or *CENPC1* nucleic acid may also be cloned into an expression vector in an antisense orientation. Regulatory sequences (e.g., viral promoters and/or enhancers) operatively linked to a *DLG1, KIAA0783, DPF3* or *CENPC1* nucleic acid cloned in the antisense orientation can be chosen for directing constitutive, tissue specific or cell type specific expression of antisense RNA in a variety of cell types. Antisense expression vectors can be in the form of a recombinant plasmid, phagemid or attenuated virus. For a discussion of the regulation of gene expression using antisense genes see Weintraub et al., Antisense RNA as a molecular tool for genetic analysis, Reviews - Trends in Genetics, Vol. 1(1) (1986).

[0072] Also provided herein are host cells that include a *DLG1, KIAA0783, DPF3* or *CENPC1* nucleic acid within a recombinant expression vector or *DLG1, KIAA0783, DPF3* or *CENPC1* nucleic acid sequence fragments which allow it to homologously recombine into a specific site of the host cell genome. The terms "host cell" and "recombinant host cell" are used interchangeably herein. Such terms refer not only to the particular subject cell but rather also to the progeny or potential progeny of such a cell. Because certain modifications may occur in succeeding generations due to either mutation or environmental influences, such progeny may not, in fact, be identical to the parent cell, but are still included within the scope of the term as used herein. A host cell can be any prokaryotic or eukaryotic cell. For example, a *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide can be expressed in bacterial

cells such as E. coli, insect cells, yeast or mammalian cells (such as Chinese hamster ovary cells (CHO) or COS cells). Other suitable host cells are known to those skilled in the art.

[0073] Vectors can be introduced into host cells via conventional transformation or transfection techniques. As used herein, the terms "transformation" and "transfection" are intended to refer to a variety of art-recognized techniques for introducing foreign nucleic acid (e.g., DNA) into a host cell, including calcium phosphate or calcium chloride co-precipitation, transduction/infection, DEAE-dextran-mediated transfection, lipofection, or electroporation.

[0074] A host cell provided herein can be used to produce (i.e., express) a *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide. Accordingly, further provided are methods for producing a *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide using the host cells described herein. In one embodiment, the method includes culturing host cells into which a recombinant expression vector encoding a *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide has been introduced in a suitable medium such that a *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide is produced. In another embodiment, the method further includes isolating a *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide from the medium or the host cell.

[0075] Also provided are cells or purified preparations of cells which include a *DLG1, KIAA0783, DPF3* or *CENPC1* transgene, or which otherwise misexpress *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide. Cell preparations can consist of human or non-human cells, e.g., rodent cells, e.g., mouse or rat cells, rabbit cells, or pig cells. In certain embodiments, the cell or cells include a *DLG1, KIAA0783, DPF3* or *CENPC1* transgene (e.g., a heterologous form of a *DLG1, KIAA0783, DPF3* or *CENPC1* such as a human gene expressed in non-human cells). The *DLG1, KIAA0783, DPF3* or *CENPC1* transgene can be misexpressed, e.g., overexpressed or underexpressed. In other embodiments, the cell or cells include a gene which misexpress an endogenous *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide (e.g., expression of a gene is disrupted, also known as a knockout). Such cells can serve as a model for studying disorders which are related to mutated or mis-expressed *DLG1, KIAA0783, DPF3* or *CENPC1* alleles or for use in drug screening. Also provided are human cells (e.g., a hematopoietic stem cells) transformed with a *DLG1, KIAA0783, DPF3* or *CENPC1* nucleic acid.

[0076] Also provided are cells or a purified preparation thereof (e.g., human cells) in which an endogenous *DLG1, KIAA0783, DPF3* or *CENPC1* nucleic acid is under the control of a regulatory sequence that does not normally control the expression of the endogenous *DLG1, KIAA0783, DPF3* or *CENPC1* gene. The expression characteristics of an endogenous gene within a cell (e.g., a cell line or microorganism) can be modified by inserting a heterologous DNA regulatory element into the genome of the cell such that the inserted regulatory element is operably linked to the endogenous *DLG1, KIAA0783, DPF3* or *CENPC1* gene. For example, an endogenous *DLG1, KIAA0783, DPF3* or *CENPC1* gene (e.g., a gene which is "transcriptionally silent," not normally expressed, or expressed only at very low levels)

may be activated by inserting a regulatory element which is capable of promoting the expression of a normally expressed gene product in that cell. Techniques such as targeted homologous recombinations, can be used to insert the heterologous DNA as described in, e.g., Chappel, US 5,272,071; WO 91/06667, published on May 16, 1991.

Transgenic Animals

[0077] Non-human transgenic animals that express a heterologous *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide (e.g., expressed from a *DLG1, KIAA0783, DPF3* or *CENPC1* nucleic acid isolated from another organism) can be generated. Such animals are useful for studying the function and/or activity of a *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide and for identifying and/or evaluating modulators of *DLG1, KIAA0783, DPF3* or *CENPC1* nucleic acid and *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide activity. As used herein, a "transgenic animal" is a non-human animal such as a mammal (e.g., a non-human primate such as chimpanzee, baboon, or macaque; an ungulate such as an equine, bovine, or caprine; or a rodent such as a rat, a mouse, or an Israeli sand rat), a bird (e.g., a chicken or a turkey), an amphibian (e.g., a frog, salamander, or newt), or an insect (e.g., Drosophila melanogaster), in which one or more of the cells of the animal includes a *DLG1, KIAA0783, DPF3* or *CENPC1* transgene. A transgene is exogenous DNA or a rearrangement (e.g., a deletion of endogenous chromosomal DNA) that is often integrated into or occurs in the genome of cells in a transgenic animal. A transgene can direct expression of an encoded gene product in one or more cell types or tissues of the transgenic animal, and other transgenes can reduce expression (e.g., a knockout). Thus, a transgenic animal can be one in which an endogenous *DLG1, KIAA0783, DPF3* or *CENPC1* gene has been altered by homologous recombination between the endogenous gene and an exogenous DNA molecule introduced into a cell of the animal (e.g., an embryonic cell of the animal) prior to development of the animal.

[0078] Intronic sequences and polyadenylation signals can also be included in the transgene to increase expression efficiency of the transgene. One or more tissue-specific regulatory sequences can be operably linked to a *DLG1, KIAA0783, DPF3* or *CENPC1* transgene to direct expression of a *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide to particular cells. A transgenic founder animal can be identified based upon the presence of a *DLG1, KIAA0783, DPF3* or *CENPC1* transgene in its genome and/or expression of *DLG1, KIAA0783, DPF3* or *CENPC1* mRNA in tissues or cells of the animals. A transgenic founder animal can then be used to breed additional animals carrying the transgene. Moreover, transgenic animals carrying a transgene encoding a *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide can further be bred to other transgenic animals carrying other transgenes.

[0079] *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptides can be expressed in transgenic animals or plants by introducing, for example, a nucleic acid encoding the polypeptide into the genome of an animal. In certain embodiments the nucleic acid is placed under the control of a tissue specific promoter, e.g., a milk or egg specific promoter, and recovered from the milk or eggs produced by the animal. Also included is a population of cells from a transgenic animal.

DLG1, KIAA0783, DPF3 and *CENPC1* Polypeptides

[0080] Featured herein are isolated *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptides, which include polypeptides having amino acid sequences set forth in SEQ ID NO: 9-12, and substantially identical polypeptides thereof. Such polypeptides sometimes are proteins or peptides. A *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide is a polypeptide encoded by a *DLG1, KIAA0783, DPF3* or *CENPC1* nucleic acid, where one nucleic acid can encode one or more different polypeptides. An "isolated" or "purified" polypeptide or protein is substantially free of cellular material or other contaminating proteins from the cell or tissue source from which the protein is derived, or substantially free from chemical precursors or other chemicals when chemically synthesized. In one embodiment, the language "substantially free" means preparation of a *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide or *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide variant having less than about 30%, 20%, 10% and sometimes 5% (by dry weight), of non-*DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide (also referred to herein as a "contaminating protein"), or of chemical precursors or non-*DLG1, KIAA0783, DPF3* or *CENPC1* chemicals. When the *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide or a biologically active portion thereof is recombinantly produced, it is also often substantially free of culture medium, specifically, where culture medium represents less than about 20%, sometimes less than about 10%, and often less than about 5% of the volume of the polypeptide preparation. Isolated or purified *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide preparations are sometimes 0.01 milligrams or more or 0.1 milligrams or more, and often 1.0 milligrams or more and 10 milligrams or more in dry weight. In specific embodiments, a *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide comprises a glutamine at amino acid position 278 in SEQ ID NO: 9 or a glycine at amino acid position 389 in SEQ ID NO: 12.

[0081] In another aspect, featured herein are *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptides and biologically active or antigenic fragments thereof that are useful as reagents or targets in assays applicable to prevention, treatment or diagnosis of breast cancer. In another embodiment, provided herein are *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptides having a *DLG1, KIAA0783, DPF3* or *CENPC1* activity or activities.

[0082] Further included herein are *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide fragments. The polypeptide fragment may be a domain or part of a domain of a *DLG1, KIAA0783, DPF3* or

CENPC1 polypeptide. The polypeptide fragment is often 50 or fewer, 100 or fewer, or 200 or fewer amino acids in length, and is sometimes 300, 400, 500, 600, 700, or 900 or fewer amino acids in length. In certain embodiments, the polypeptide fragment comprises, consists essentially of, or consists of, at least 6 consecutive amino acids and not more than 1211 consecutive amino acids of SEQ ID NO: 9-12, or the polypeptide fragment comprises, consists essentially of, or consists of, at least 6 consecutive amino acids and not more than 543 consecutive amino acids of SEQ ID NO: 9-12.

[0083] *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptides described herein can be used as immunogens to produce anti-*DLG1, KIAA0783, DPF3* or *CENPC1* antibodies in a subject, to purify *DLG1, KIAA0783, DPF3* or *CENPC1* ligands or binding partners, and in screening assays to identify molecules which inhibit or enhance the interaction of *DLG1, KIAA0783, DPF3* or *CENPC1* with a *DLG1, KIAA0783, DPF3* or *CENPC1* substrate. Full-length *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptides and polynucleotides encoding the same may be specifically substituted for a *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide fragment or polynucleotide encoding the same in any embodiment described herein.

[0084] Substantially identical polypeptides may depart from the amino acid sequences set forth in SEQ ID NO: 9-12 in different manners. For example, conservative amino acid modifications may be introduced at one or more positions in the amino acid sequences of SEQ ID NO: 9-12. A "conservative amino acid substitution" is one in which the amino acid is replaced by another amino acid having a similar structure and/or chemical function. Families of amino acid residues having similar structures and functions are well known. These families include amino acids with basic side chains (e.g., lysine, arginine, histidine), acidic side chains (e.g., aspartic acid, glutamic acid), uncharged polar side chains (e.g., glycine, asparagine, glutamine, serine, threonine, tyrosine, cysteine), nonpolar side chains (e.g., alanine, valine, leucine, isoleucine, proline, phenylalanine, methionine, tryptophan), beta-branched side chains (e.g., threonine, valine, isoleucine) and aromatic side chains (e.g., tyrosine, phenylalanine, tryptophan, histidine). Also, essential and non-essential amino acids may be replaced. A "non-essential" amino acid is one that can be altered without abolishing or substantially altering the biological function of a *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide, whereas altering an "essential" amino acid abolishes or substantially alters the biological function of a *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide. Amino acids that are conserved among *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptides are typically essential amino acids.

[0085] Also, *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptides and polypeptide variants may exist as chimeric or fusion polypeptides. As used herein, a *DLG1, KIAA0783, DPF3* or *CENPC1* "chimeric polypeptide" or "fusion polypeptide" includes a *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide linked to a non-*DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide. A "non-*DLG1, KIAA0783, DPF3* or

CENPC1 polypeptide" refers to a polypeptide having an amino acid sequence corresponding to a polypeptide which is not substantially identical to the *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide, which includes, for example, a polypeptide that is different from the *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide and derived from the same or a different organism. The *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide in the fusion polypeptide can correspond to an entire or nearly entire *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide or a fragment thereof. The non-*DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide can be fused to the N-terminus or C-terminus of the *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide.

[0086] Fusion polypeptides can include a moiety having high affinity for a ligand. For example, the fusion polypeptide can be a GST-*DLG1, KIAA0783, DPF3* or *CENPC1* fusion polypeptide in which the *DLG1, KIAA0783, DPF3* or *CENPC1* sequences are fused to the C-terminus of the GST sequences, or a polyhistidine-*DLG1, KIAA0783, DPF3* or *CENPC1* fusion polypeptide in which the *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide is fused at the N- or C-terminus to a string of histidine residues. Such fusion polypeptides can facilitate purification of recombinant *DLG1, KIAA0783, DPF3* or *CENPC1*. Expression vectors are commercially available that already encode a fusion moiety (e.g., a GST polypeptide), and a *DLG1, KIAA0783, DPF3* or *CENPC1* nucleic acid can be cloned into an expression vector such that the fusion moiety is linked in-frame to the *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide. Further, the fusion polypeptide can be a *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide containing a heterologous signal sequence at its N-terminus. In certain host cells (e.g., mammalian host cells), expression, secretion, cellular internalization, and cellular localization of a *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide can be increased through use of a heterologous signal sequence. Fusion polypeptides can also include all or a part of a serum polypeptide (e.g., an IgG constant region or human serum albumin).

[0087] *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptides or fragments thereof can be incorporated into pharmaceutical compositions and administered to a subject in vivo. Administration of these *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptides can be used to affect the bioavailability of a *DLG1, KIAA0783, DPF3* or *CENPC1* substrate and may effectively increase or decrease *DLG1, KIAA0783, DPF3* or *CENPC1* biological activity in a cell or effectively supplement dysfunctional or hyperactive *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide. *DLG1, KIAA0783, DPF3* or *CENPC1* fusion polypeptides may be useful therapeutically for the treatment of disorders caused by, for example, (i) aberrant modification or mutation of a gene encoding a *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide; (ii) mis-regulation of the *DLG1, KIAA0783, DPF3* or *CENPC1* gene; and (iii) aberrant post-translational modification of a *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide. Also, *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptides can be used as immunogens to produce anti-*DLG1,*

KIAA0783, DPF3 or *CENPC1* antibodies in a subject, to purify *DLG1, KIAA0783, DPF3* or *CENPC1* ligands or binding partners, and in screening assays to identify molecules which inhibit or enhance the interaction of *DLG1, KIAA0783, DPF3* or *CENPC1* with a *DLG1, KIAA0783, DPF3* or *CENPC1* substrate. Preferably, said *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptides are used in screening assays to identify molecules which inhibit the interaction of *DLG1, KIAA0783, DPF3* or *CENPC1*.

[0088] In addition, polypeptides can be chemically synthesized using techniques known in the art (See, *e.g.*, Creighton, 1983 Proteins. New York, N.Y.: W. H. Freeman and Company; and Hunkapiller *et al.*, (1984) Nature July 12 -18;310(5973):105-11). For example, a relative short polypeptide fragment can be synthesized by use of a peptide synthesizer. Furthermore, if desired, non-classical amino acids or chemical amino acid analogs can be introduced as a substitution or addition into the fragment sequence. Non-classical amino acids include, but are not limited to, to the D-isomers of the common amino acids, 2,4-diaminobutyric acid, a-amino isobutyric acid, 4-aminobutyric acid, Abu, 2-amino butyric acid, g-Abu, e-Ahx, 6-amino hexanoic acid, Aib, 2-amino isobutyric acid, 3-amino propionic acid, ornithine, norleucine, norvaline, hydroxyproline, sarcosine, citrulline, homocitrulline, cysteic acid, t-butylglycine, t-butylalanine, phenylglycine, cyclohexylalanine, b-alanine, fluoroamino acids, designer amino acids such as b-methyl amino acids, Ca-methyl amino acids, Na-methyl amino acids, and amino acid analogs in general. Furthermore, the amino acid can be D (dextrorotary) or L (levorotary).

[0089] Also included are polypeptide fragments which are differentially modified during or after translation, *e.g.*, by glycosylation, acetylation, phosphorylation, amidation, derivatization by known protecting/blocking groups, proteolytic cleavage, linkage to an antibody molecule or other cellular ligand, and the like. Any of numerous chemical modifications may be carried out by known techniques, including but not limited, to specific chemical cleavage by cyanogen bromide, trypsin, chymotrypsin, papain, V8 protease, $NaBH_4$; acetylation, formylation, oxidation, reduction; metabolic synthesis in the presence of tunicamycin; and the like.

[0090] Additional post-translational modifications include, for example, N-linked or O-linked carbohydrate chains, processing of N-terminal or C-terminal ends), attachment of chemical moieties to the amino acid backbone, chemical modifications of N-linked or O-linked carbohydrate chains, and addition or deletion of an N-terminal methionine residue as a result of prokaryotic host cell expression. The polypeptide fragments may also be modified with a detectable label, such as an enzymatic, fluorescent, isotopic or affinity label to allow for detection and isolation of the polypeptide.

[0091] Also provided are chemically modified polypeptide derivatives that may provide additional advantages such as increased solubility, stability and circulating time of the polypeptide, or decreased immunogenicity. See U.S. Pat. No: 4,179,337. The chemical moieties for derivitization may be selected from water soluble polymers such as polyethylene glycol, ethylene glycol/propylene glycol copolymers,

carboxymethylcellulose, dextran, polyvinyl alcohol and the like. The polypeptides may be modified at random positions within the molecule, or at predetermined positions within the molecule and may include one, two, three or more attached chemical moieties.

[0092] The polymer may be of any molecular weight, and may be branched or unbranched. For polyethylene glycol, the molecular weight is between about 1 kDa and about 100 kDa (the term "about" indicating that in preparations of polyethylene glycol, some molecules will weigh more, some less, than the stated molecular weight) for ease in handling and manufacturing. Other sizes may be used, depending on the desired therapeutic profile (*e.g.*, the duration of sustained release desired, the effects, if any on biological activity, the ease in handling, the degree or lack of antigenicity and other known effects of the polyethylene glycol to a therapeutic protein or analog).

[0093] The polyethylene glycol molecules (or other chemical moieties) should be attached to the polypeptide with consideration of effects on functional or antigenic domains of the polypeptide. There are a number of attachment methods available to those skilled in the art, *e.g.*, EP 0 401 384, herein incorporated by reference (coupling PEG to G-CSF), see also Malik *et al.* (1992) Exp Hematol. September;20(8):1028-35, reporting pegylation of GM-CSF using tresyl chloride). For example, polyethylene glycol may be covalently bound through amino acid residues via a reactive group, such as, a free amino or carboxyl group. Reactive groups are those to which an activated polyethylene glycol molecule may be bound. The amino acid residues having a free amino group may include lysine residues and the N-terminal amino acid residues; those having a free carboxyl group may include aspartic acid residues, glutamic acid residues and the C-terminal amino acid residue. Sulfhydryl groups may also be used as a reactive group for attaching the polyethylene glycol molecules. A polymer sometimes is attached at an amino group, such as attachment at the N-terminus or lysine group.

[0094] One may specifically desire proteins chemically modified at the N-terminus. Using polyethylene glycol as an illustration of the present composition, one may select from a variety of polyethylene glycol molecules (by molecular weight, branching, and the like), the proportion of polyethylene glycol molecules to protein (polypeptide) molecules in the reaction mix, the type of pegylation reaction to be performed, and the method of obtaining the selected N-terminally pegylated protein. The method of obtaining the N-terminally pegylated preparation (i.e., separating this moiety from other monopegylated moieties if necessary) may be by purification of the N-terminally pegylated material from a population of pegylated protein molecules. Selective proteins chemically modified at the N-terminus may be accomplished by reductive alkylation, which exploits differential reactivity of different types of primary amino groups (lysine versus the N-terminal) available for derivatization in a particular protein. Under the appropriate reaction conditions, substantially selective derivatization of the protein at the N-terminus with a carbonyl group containing polymer is achieved.

Substantially Identical Nucleic Acids and Polypeptides

[0095] Nucleotide sequences and polypeptide sequences that are substantially identical to a *DLG1, KIAA0783, DPF3* or *CENPC1* nucleotide sequence and the *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide sequences encoded by those nucleotide sequences are included herein. The term "substantially identical" as used herein refers to two or more nucleic acids or polypeptides sharing one or more identical nucleotide sequences or polypeptide sequences, respectively. Included are nucleotide sequences or polypeptide sequences that are 55% or more, 60% or more, 65% or more, 70% or more, 75% or more, 80% or more, 85% or more, 90% or more, 95% or more (each often within a 1%, 2%, 3% or 4% variability) or more identical to the nucleotide sequences in SEQ ID NO: 1-8 or the encoded *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide amino acid sequences. One test for determining whether two nucleic acids are substantially identical is to determine the percent of identical nucleotide sequences or polypeptide sequences shared between the nucleic acids or polypeptides.

[0096] Calculations of sequence identity are often performed as follows. Sequences are aligned for optimal comparison purposes (*e.g.*, gaps can be introduced in one or both of a first and a second amino acid or nucleic acid sequence for optimal alignment and non-homologous sequences can be disregarded for comparison purposes). The length of a reference sequence aligned for comparison purposes is sometimes 30% or more, 40% or more, 50% or more, often 60% or more, and more often 70% or more, 80% or more, 90% or more, 90% or more, or 100% of the length of the reference sequence. The nucleotides or amino acids at corresponding nucleotide or polypeptide positions, respectively, are then compared among the two sequences. When a position in the first sequence is occupied by the same nucleotide or amino acid as the corresponding position in the second sequence, the nucleotides or amino acids are deemed to be identical at that position. The percent identity between the two sequences is a function of the number of identical positions shared by the sequences, taking into account the number of gaps, and the length of each gap, introduced for optimal alignment of the two sequences.

[0097] Comparison of sequences and determination of percent identity between two sequences can be accomplished using a mathematical algorithm. Percent identity between two amino acid or nucleotide sequences can be determined using the algorithm of Meyers & Miller, *CABIOS 4:* 11-17 (1989), which has been incorporated into the ALIGN program (version 2.0), using a PAM120 weight residue table, a gap length penalty of 12 and a gap penalty of 4. Also, percent identity between two amino acid sequences can be determined using the Needleman & Wunsch, *J. Mol. Biol. 48:* 444-453 (1970) algorithm which has been incorporated into the GAP program in the GCG software package (available at the http address www.gcg.com), using either a Blossum 62 matrix or a PAM250 matrix, and a gap weight of 16, 14, 12, 10, 8, 6, or 4 and a length weight of 1, 2, 3, 4, 5, or 6. Percent identity between two nucleotide sequences can be determined using the GAP program in the GCG software package (available

at http address www.gcg.com), using a NWSgapdna.CMP matrix and a gap weight of 40, 50, 60, 70, or 80 and a length weight of 1, 2, 3, 4, 5, or 6. A set of parameters often used is a Blossum 62 scoring matrix with a gap open penalty of 12, a gap extend penalty of 4, and a frameshift gap penalty of 5.

[0098] Another manner for determining if two nucleic acids are substantially identical is to assess whether a polynucleotide homologous to one nucleic acid will hybridize to the other nucleic acid under stringent conditions. As use herein, the term "stringent conditions" refers to conditions for hybridization and washing. Stringent conditions are known to those skilled in the art and can be found in *Current Protocols in Molecular Biology*, John Wiley & Sons, N.Y. , 6.3.1-6.3.6 (1989). Aqueous and non-aqueous methods are described in that reference and either can be used. An example of stringent hybridization conditions is hybridization in 6X sodium chloride/sodium citrate (SSC) at about 45°C, followed by one or more washes in 0.2X SSC, 0.1% SDS at 50°C. Another example of stringent hybridization conditions are hybridization in 6X sodium chloride/sodium citrate (SSC) at about 45°C, followed by one or more washes in 0.2X SSC, 0.1% SDS at 55°C. A further example of stringent hybridization conditions is hybridization in 6X sodium chloride/sodium citrate (SSC) at about 45°C, followed by one or more washes in 0.2X SSC, 0.1% SDS at 60°C. Often, stringent hybridization conditions are hybridization in 6X sodium chloride/sodium citrate (SSC) at about 45°C, followed by one or more washes in 0.2X SSC, 0.1% SDS at 65°C. More often, stringency conditions are 0.5M sodium phosphate, 7% SDS at 65°C, followed by one or more washes at 0.2X SSC, 1% SDS at 65°C.

[0099] An example of a substantially identical nucleotide sequence to a *DLG1, KIAA0783, DPF3* or *CENPC1* nucleotide sequence is one that has a different nucleotide sequence but still encodes the same polypeptide sequence encoded by the *DLG1, KIAA0783, DPF3* or *CENPC1* nucleotide sequence. Another example is a nucleotide sequence that encodes a polypeptide having a polypeptide sequence that is more than 70% or more identical to, sometimes 75% or more, 80% or more, or 85% or more identical to, and often 90% or more and 95% or more identical to a polypeptide sequence encoded by a *DLG1, KIAA0783, DPF3* or *CENPC1* nucleotide sequence.

[0100] *DLG1, KIAA0783, DPF3* or *CENPC1* nucleotide sequences and *DLG1, KIAA0783, DPF3* or *CENPC1* amino acid sequences can be used as "query sequences" to perform a search against public databases to identify other family members or related sequences, for example. Such searches can be performed using the NBLAST and XBLAST programs (version 2.0) of Altschul *et al., J. Mol. Biol. 215:* 403-10 (1990). BLAST nucleotide searches can be performed with the NBLAST program, score = 100, wordlength = 12 to obtain nucleotide sequences homologous to nucleotide sequences from SEQ ID NO: 1-8. BLAST polypeptide searches can be performed with the XBLAST program, score = 50, wordlength = 3 to obtain amino acid sequences homologous to polypeptides encoded by a *DLG1, KIAA0783, DPF3* or *CENPC1* nucleotide sequence. To obtain gapped alignments for comparison purposes, Gapped

BLAST can be utilized as described in Altschul *et al., Nucleic Acids Res. 25(17):* 3389-3402 (1997). When utilizing BLAST and Gapped BLAST programs, default parameters of the respective programs (*e.g.*, XBLAST and NBLAST) can be used (*see* the http address www.ncbi.nlm.nih.gov).

[0101] A nucleic acid that is substantially identical to a *DLG1, KIAA0783, DPF3* or *CENPC1* nucleotide sequence may include polymorphic sites at positions equivalent to those described herein when the sequences are aligned. For example, using the alignment procedures described herein, SNPs in a sequence substantially identical to a sequence in SEQ ID NO: 1-8 can be identified at nucleotide positions that match (*i.e.*, align) with nucleotides at SNP positions in the nucleotide sequence of SEQ ID NO: 1-8. Also, where a polymorphic variation results in an insertion or deletion, insertion or deletion of a nucleotide sequence from a reference sequence can change the relative positions of other polymorphic sites in the nucleotide sequence.

[0102] Substantially identical nucleotide and polypeptide sequences include those that are naturally occurring, such as allelic variants (same locus), splice variants, homologs (different locus), and orthologs (different organism) or can be non-naturally occurring. Non-naturally occurring variants can be generated by mutagenesis techniques, including those applied to polynucleotides, cells, or organisms. The variants can contain nucleotide substitutions, deletions, inversions and insertions. Variation can occur in either or both the coding and non-coding regions. The variations can produce both conservative and non-conservative amino acid substitutions (as compared in the encoded product). Orthologs, homologs, allelic variants, and splice variants can be identified using methods known in the art. These variants normally comprise a nucleotide sequence encoding a polypeptide that is 50% or more, about 55% or more, often about 70-75% or more, more often about 80-85% or more, and typically about 90-95% or more identical to the amino acid sequences of target polypeptides or a fragment thereof. Such nucleic acid molecules readily can be identified as being able to hybridize under stringent conditions to a nucleotide sequence in SEQ ID NO: 1-8 or a fragment thereof. Nucleic acid molecules corresponding to orthologs, homologs, and allelic variants of a nucleotide sequence in SEQ ID NO: 1-8 can be identified by mapping the sequence to the same chromosome or locus as the nucleotide sequence in SEQ ID NO: 1-8.

[0103] Also, substantially identical nucleotide sequences may include codons that are altered with respect to the naturally occurring sequence for enhancing expression of a target polypeptide in a particular expression system. For example, the nucleic acid can be one in which one or more codons are altered, and often 10% or more or 20% or more of the codons are altered for optimized expression in bacteria (*e.g.*, *E. coli.*), yeast (*e.g.*, *S. cervesiae*), human (*e.g.*, 293 cells), insect, or rodent (*e.g.*, hamster) cells.

Methods for Identifying Subjects at Risk of Breast Cancer and Breast Cancer Risk in a Subject

[0104] Methods for prognosing and diagnosing breast cancer in subjects are provided herein. These methods include detecting the presence or absence of one or more polymorphic variations associated with breast cancer in a nucleotide sequence set forth in SEQ ID NO: 1-4, or substantially identical sequence thereof, in a sample from a subject, where the presence of a polymorphic variant is indicative of a risk of breast cancer.

[0105] Thus, featured herein is a method for detecting a subject at risk of breast cancer or the risk of breast cancer in a subject, which comprises detecting the presence or absence of a polymorphic variation associated with breast cancer at a polymorphic site in a nucleotide sequence set forth in SEQ ID NO: 1-4 in a nucleic acid sample from a subject, where the nucleotide sequence comprises a polynucleotide sequence selected from the group consisting of: (a) a nucleotide sequence set forth in SEQ ID NO: 1-4; (b) a nucleotide sequence which encodes a polypeptide having an amino acid sequence encoded by a nucleotide sequence in SEQ ID NO: 1-4; (c) a nucleotide sequence which encodes a polypeptide that is 90% or more identical to an amino acid sequence encoded by a nucleotide sequence in SEQ ID NO: 1-4 or a nucleotide sequence about 90% or more identical to the nucleotide sequence set forth in SEQ ID NO: 1-4; and (d) a fragment of a nucleotide sequence of (a), (b), or (c), often a fragment that includes a polymorphic site associated with breast cancer; whereby the presence of the polymorphic variation is indicative of a risk of breast cancer in the subject. In certain embodiments, determining the presence of a combination of two or more polymorphic variants associated with breast cancer in one or more nucleotide sequences of the sample is determined to identify a subject at risk of breast cancer and/or risk of breast cancer.

[0106] A risk of developing aggressive forms of breast cancer likely to metastasize or invade surrounding tissues (e.g., Stage IIIA, IIIB, and IV breast cancers), and subjects at risk of developing aggressive forms of breast cancer also may be identified by the methods described herein. These methods include collecting phenotype information from subjects having breast cancer, which includes the stage of progression of the breast cancer, and performing a secondary phenotype analysis to detect the presence or absence of one or more polymorphic variations associated with a particular stage form of breast cancer. Thus, detecting the presence or absence of one or more polymorphic variations in a *DLG1, KIAA0783, DPF3* or *CENPC1* nucleotide sequence associated with a late stage form of breast cancer often is prognostic and/or diagnostic of an aggressive form of the cancer.

[0107] Results from prognostic tests may be combined with other test results to diagnose breast cancer. For example, prognostic results may be gathered, a patient sample may be ordered based on a determined predisposition to breast cancer, the patient sample is analyzed, and the results of the analysis may be utilized to diagnose breast cancer. Also breast cancer diagnostic methods can be developed from

studies used to generate prognostic/diagnostic methods in which populations are stratified into subpopulations having different progressions of breast cancer. In another embodiment, prognostic results may be gathered; a patient's risk factors for developing breast cancer analyzed (*e.g.*, age, race, family history, age of first menstrual cycle, age at birth of first child); and a patient sample may be ordered based on a determined predisposition to breast cancer. In an alternative embodiment, the results from predisposition analyses described herein may be combined with other test results indicative of breast cancer, which were previously, concurrently, or subsequently gathered with respect to the predisposition testing. In these embodiments, the combination of the prognostic test results with other test results can be probative of breast cancer, and the combination can be utilized as a breast cancer diagnostic. The results of any test indicative of breast cancer known in the art may be combined with the methods described herein. Examples of such tests are mammography (*e.g.*, a more frequent and/or earlier mammography regimen may be prescribed); breast biopsy and optionally a biopsy from another tissue; breast ultrasound and optionally an ultrasound analysis of another tissue; breast magnetic resonance imaging (MRI) and optionally an MRI analysis of another tissue; electrical impedance (T-scan) analysis of breast and optionally of another tissue; ductal lavage; nuclear medicine analysis (*e.g.*, scintimammography); *BRCA1* and/or *BRCA2* sequence analysis results; and thermal imaging of the breast and optionally of another tissue. Testing may be performed on tissue other than breast to diagnose the occurrence of metastasis (*e.g.*, testing of the lymph node).

[0108] Risk of breast cancer sometimes is expressed as a probability, such as an odds ratio, percentage, or risk factor. The risk is based upon the presence or absence of one or more polymorphic variants described herein, and also may be based in part upon phenotypic traits of the individual being tested. Methods for calculating predispositions based upon patient data are well known (*see, e.g.*, Agresti, *Categorical Data Analysis*, 2nd Ed. 2002. Wiley). Allelotyping and genotyping analyses may be carried out in populations other than those exemplified herein to enhance the predictive power of the prognostic method. These further analyses are executed in view of the exemplified procedures described herein, and may be based upon the same polymorphic variations or additional polymorphic variations. Risk determinations for breast cancer are useful in a variety of applications. In one embodiment, breast cancer risk determinations are used by clinicians to direct appropriate detection, preventative and treatment procedures to subjects who most require these. In another embodiment, breast cancer risk determinations are used by health insurers for preparing actuarial tables and for calculating insurance premiums.

[0109] The nucleic acid sample typically is isolated from a biological sample obtained from a subject. For example, nucleic acid can be isolated from blood, saliva, sputum, urine, cell scrapings, and biopsy tissue. The nucleic acid sample can be isolated from a biological sample using standard

techniques, such as the technique described in Example 2. As used herein, the term "subject" refers primarily to humans but also refers to other mammals such as dogs, cats, and ungulates (*e.g.*, cattle, sheep, and swine). Subjects also include avians (*e.g.*, chickens and turkeys), reptiles, and fish (*e.g.*, salmon), as embodiments described herein can be adapted to nucleic acid samples isolated from any of these organisms. The nucleic acid sample may be isolated from the subject and then directly utilized in a method for determining the presence of a polymorphic variant, or alternatively, the sample may be isolated and then stored (*e.g.*, frozen) for a period of time before being subjected to analysis.

[0110] The presence or absence of a polymorphic variant is determined using one or both chromosomal complements represented in the nucleic acid sample. Determining the presence or absence of a polymorphic variant in both chromosomal complements represented in a nucleic acid sample from a subject having a copy of each chromosome is useful for determining the zygosity of an individual for the polymorphic variant (*i.e.*, whether the individual is homozygous or heterozygous for the polymorphic variant). Any oligonucleotide-based diagnostic may be utilized to determine whether a sample includes the presence or absence of a polymorphic variant in a sample. For example, primer extension methods, ligase sequence determination methods (*e.g.*, U.S. Pat. Nos. 5,679,524 and 5,952,174, and WO 01/27326), mismatch sequence determination methods (*e.g.*, U.S. Pat. Nos. 5,851,770; 5,958,692; 6,110,684; and 6,183,958), microarray sequence determination methods, restriction fragment length polymorphism (RFLP), single strand conformation polymorphism detection (SSCP) (*e.g.*, U.S. Pat. Nos. 5,891,625 and 6,013,499), PCR-based assays (*e.g.*, TAQMAN® PCR System (Applied Biosystems)), and nucleotide sequencing methods may be used.

[0111] Oligonucleotide extension methods typically involve providing a pair of oligonucleotide primers in a polymerase chain reaction (PCR) or in other nucleic acid amplification methods for the purpose of amplifying a region from the nucleic acid sample that comprises the polymorphic variation. One oligonucleotide primer is complementary to a region 3' of the polymorphism and the other is complementary to a region 5' of the polymorphism. A PCR primer pair may be used in methods disclosed in U.S. Pat. Nos. 4,683,195; 4,683,202, 4,965,188; 5,656,493; 5,998,143; 6,140,054; WO 01/27327; and WO 01/27329 for example. PCR primer pairs may also be used in any commercially available machines that perform PCR, such as any of the GENEAMP® Systems available from Applied Biosystems. Also, those of ordinary skill in the art will be able to design oligonucleotide primers based upon a nucleotide sequence set forth in SEQ ID NO: 1-4 without undue experimentation using knowledge readily available in the art.

[0112] Also provided is an extension oligonucleotide that hybridizes to the amplified fragment adjacent to the polymorphic variation. As used herein, the term "adjacent" refers to the 3' end of the extension oligonucleotide being often 1 nucleotide from the 5' end of the polymorphic site, and

sometimes 2, 3, 4, 5, 6, 7, 8, 9, or 10 nucleotides from the 5' end of the polymorphic site, in the nucleic acid when the extension oligonucleotide is hybridized to the nucleic acid. The extension oligonucleotide then is extended by one or more nucleotides, and the number and/or type of nucleotides that are added to the extension oligonucleotide determine whether the polymorphic variant is present. Oligonucleotide extension methods are disclosed, for example, in U.S. Pat. Nos. 4,656,127; 4,851,331; 5,679,524; 5,834,189; 5,876,934; 5,908,755; 5,912,118; 5,976,802; 5,981,186; 6,004,744; 6,013,431; 6,017,702; 6,046,005; 6,087,095; 6,210,891; and WO 01/20039. Oligonucleotide extension methods using mass spectrometry are described, for example, in U.S. Pat. Nos. 5,547,835; 5,605,798; 5,691,141; 5,849,542; 5,869,242; 5,928,906; 6,043,031; and 6,194,144, and a method often utilized is described herein in Example 2. Multiple extension oligonucleotides may be utilized in one reaction, which is referred to herein as "multiplexing."

[0113] A microarray can be utilized for determining whether a polymorphic variant is present or absent in a nucleic acid sample. A microarray may include any oligonucleotides described herein, and methods for making and using oligonucleotide microarrays suitable for diagnostic use are disclosed in U.S. Pat. Nos. 5,492,806; 5,525,464; 5,589,330; 5,695,940; 5,849,483; 6,018,041; 6,045,996; 6,136,541; 6,142,681; 6,156,501; 6,197,506; 6,223,127; 6,225,625; 6,229,911; 6,239,273; WO 00/52625; WO 01/25485; and WO 01/29259. The microarray typically comprises a solid support and the oligonucleotides may be linked to this solid support by covalent bonds or by non-covalent interactions. The oligonucleotides may also be linked to the solid support directly or by a spacer molecule. A microarray may comprise one or more oligonucleotides complementary to a polymorphic site set forth in SEQ ID NO: 1-4 or below.

[0114] A kit also may be utilized for determining whether a polymorphic variant is present or absent in a nucleic acid sample. A kit often comprises one or more pairs of oligonucleotide primers useful for amplifying a fragment of a *DLG1, KIAA0783, DPF3* or *CENPC1* nucleotide sequence or a substantially identical sequence thereof, where the fragment includes a polymorphic site. The kit sometimes comprises a polymerizing agent, for example, a thermostable nucleic acid polymerase such as one disclosed in U.S. Pat. Nos. 4,889,818 or 6,077,664. Also, the kit often comprises an elongation oligonucleotide that hybridizes to a *DLG1, KIAA0783, DPF3* or *CENPC1* nucleotide sequence in a nucleic acid sample adjacent to the polymorphic site. Where the kit includes an elongation oligonucleotide, it also often comprises chain elongating nucleotides, such as dATP, dTTP, dGTP, dCTP, and dITP, including analogs of dATP, dTTP, dGTP, dCTP and dITP, provided that such analogs are substrates for a thermostable nucleic acid polymerase and can be incorporated into a nucleic acid chain elongated from the extension oligonucleotide. Along with chain elongating nucleotides would be one or more chain terminating nucleotides such as ddATP, ddTTP, ddGTP, ddCTP, and the like. In an

embodiment, the kit comprises one or more oligonucleotide primer pairs, a polymerizing agent, chain elongating nucleotides, at least one elongation oligonucleotide, and one or more chain terminating nucleotides. Kits optionally include buffers, vials, microtiter plates, and instructions for use.

[0115] An individual identified as being at risk of breast cancer may be heterozygous or homozygous with respect to the allele associated with a higher risk of breast cancer. A subject homozygous for an allele associated with an increased risk of breast cancer is at a comparatively high risk of breast cancer, a subject heterozygous for an allele associated with an increased risk of breast cancer is at a comparatively intermediate risk of breast cancer, and a subject homozygous for an allele associated with a decreased risk of breast cancer is at a comparatively low risk of breast cancer. A genotype may be assessed for a complementary strand, such that the complementary nucleotide at a particular position is detected.

[0116] Also featured are methods for determining risk of breast cancer and/or identifying a subject at risk of breast cancer by contacting a polypeptide or protein encoded by a *DLG1, KIAA0783, DPF3* or *CENPC1* nucleotide sequence from a subject with an antibody that specifically binds to an epitope associated with increased risk of breast cancer in the polypeptide. In certain embodiments, the antibody specifically binds to an epitope that comprises a glutamine at amino acid position 278 in SEQ ID NO: 9 or a glycine at amino acid position 389 in SEQ ID NO: 12.

Applications of Prognostic and Diagnostic Results to Pharmacogenomic Methods

[0117] Pharmacogenomics is a discipline that involves tailoring a treatment for a subject according to the subject's genotype. For example, based upon the outcome of a prognostic test described herein, a clinician or physician may target pertinent information and preventative or therapeutic treatments to a subject who would be benefited by the information or treatment and avoid directing such information and treatments to a subject who would not be benefited (*e.g.*, the treatment has no therapeutic effect and/or the subject experiences adverse side effects). As therapeutic approaches for breast cancer continue to evolve and improve, the goal of treatments for breast cancer related disorders is to intervene even before clinical signs (*e.g.*, identification of lump in the breast) first manifest. Thus, genetic markers associated with susceptibility to breast cancer prove useful for early diagnosis, prevention and treatment of breast cancer.

[0118] The following is an example of a pharmacogenomic embodiment. A particular treatment regimen can exert a differential effect depending upon the subject's genotype. Where a candidate therapeutic exhibits a significant interaction with a major allele and a comparatively weak interaction with a minor allele (*e.g.*, an order of magnitude or greater difference in the interaction), such a therapeutic typically would not be administered to a subject genotyped as being homozygous for the

minor allele, and sometimes not administered to a subject genotyped as being heterozygous for the minor allele. In another example, where a candidate therapeutic is not significantly toxic when administered to subjects who are homozygous for a major allele but is comparatively toxic when administered to subjects heterozygous or homozygous for a minor allele, the candidate therapeutic is not typically administered to subjects who are genotyped as being heterozygous or homozygous with respect to the minor allele.

[0119] The methods described herein are applicable to pharmacogenomic methods for detecting, preventing, alleviating and/or treating breast cancer. For example, a nucleic acid sample from an individual may be subjected to a genetic test described herein. Where one or more polymorphic variations associated with increased risk of breast cancer are identified in a subject, information for detecting, preventing or treating breast cancer and/or one or more breast cancer detection, prevention and/or treatment regimens then may be directed to and/or prescribed to that subject.

[0120] In certain embodiments, a detection, prevenative and/or treatment regimen is specifically prescribed and/or administered to individuals who will most benefit from it based upon their risk of developing breast cancer assessed by the methods described herein. Thus, provided are methods for identifying a subject at risk of breast cancer and then prescribing a detection, therapeutic or preventative regimen to individuals identified as being at risk of breast cancer. Thus, certain embodiments are directed to methods for treating breast cancer in a subject, reducing risk of breast cancer in a subject, or early detection of breast cancer in a subject, which comprise: detecting the presence or absence of a polymorphic variant associated with breast cancer in a nucleotide sequence in a nucleic acid sample from a subject, where the nucleotide sequence comprises a polynucleotide sequence selected from the group consisting of: (a) a nucleotide sequence set forth in SEQ ID NO: 1-4; (b) a nucleotide sequence which encodes a polypeptide having an amino acid sequence encoded by a nucleotide sequence in SEQ ID NO: 1-4; (c) a nucleotide sequence which encodes a polypeptide that is 90% or more identical to an amino acid sequence encoded by a nucleotide sequence in SEQ ID NO: 1-4 or a nucleotide sequence about 90% or more identical to the nucleotide sequence set forth in SEQ ID NO: 1-4; and (d) a fragment of a nucleotide sequence of (a), (b), or (c), sometimes comprising a polymorphic site associated with breast cancer; and prescribing or administering a breast cancer treatment regimen, preventative regimen and/or detection regimen to a subject from whom the sample originated where the presence of one or more polymorphic variations associated with breast cancer are detected in the nucleotide sequence. In these methods, genetic results may be utilized in combination with other test results to diagnose breast cancer as described above. Other test results include but are not limited to mammography results, imaging results, biopsy results and results from *BRCA1* or *BRAC2* test results, as described above.

[0121] Detection regimens include one or more mammography procedures, a regular mammography regimen (*e.g.*, once a year, or once every six, four, three or two months); an early mammography regimen

(*e.g.*, mammography tests are performed beginning at age 25, 30, or 35); one or more biopsy procedures (*e.g.*, a regular biopsy regimen beginning at age 40); breast biopsy and biopsy from other tissue; breast ultrasound and optionally ultrasound analysis of another tissue; breast magnetic resonance imaging (MRI) and optionally MRI analysis of another tissue; electrical impedance (T-scan) analysis of breast and optionally another tissue; ductal lavage; nuclear medicine analysis (*e.g.*, scintimammography); *BRCA1* and/or *BRCA2* sequence analysis results; and/or thermal imaging of the breast and optionally another tissue.

[0122] Treatments sometimes are preventative (*e.g.*, is prescribed or administered to reduce the probability that a breast cancer associated condition arises or progresses), sometimes are therapeutic, and sometimes delay, alleviate or halt the progression of breast cancer. Any known preventative or therapeutic treatment for alleviating or preventing the occurrence of breast cancer is prescribed and/or administered. For example, certain preventative treatments often are prescribed to subjects having a predisposition to breast cancer and where the subject is not diagnosed with breast cancer or is diagnosed as having symptoms indicative of early stage breast cancer (*e.g.*, stage I). For subjects not diagnosed as having breast cancer, any preventative treatments known in the art can be prescribed and administered, which include selective hormone receptor modulators (*e.g.*, selective estrogen receptor modulators (SERMs) such as tamoxifen, reloxifene, and toremifene); compositions that prevent production of hormones (*e.g.*, aramotase inhibitors that prevent the production of estrogen in the adrenal gland, such as exemestane, letrozole, anastrozol, groserelin, and megestrol); other hormonal treatments (*e.g.*, goserelin acetate and fulvestrant); biologic response modifiers such as antibodies (*e.g.*, trastuzumab (herceptin/HER2)); surgery (*e.g.*, lumpectomy and mastectomy); drugs that delay or halt metastasis (*e.g.*, pamidronate disodium); and alternative/complementary medicine (*e.g.*, acupuncture, acupressure, moxibustion, qi gong, reiki, ayurveda, vitamins, minerals, and herbs (*e.g.*, astragalus root, burdock root, garlic, green tea, and licorice root)).

[0123] The use of breast cancer treatments are well known in the art, and include surgery, chemotherapy and/or radiation therapy. Any of the treatments may be used in combination to treat or prevent breast cancer (*e.g.*, surgery followed by radiation therapy or chemotherapy). Examples of chemotherapy combinations used to treat breast cancer include: cyclophosphamide (Cytoxan), methotrexate (Amethopterin, Mexate, Folex), and fluorouracil (Fluorouracil, 5-Fu, Adrucil), which is referred to as CMF; cyclophosphamide, doxorubicin (Adriamycin), and fluorouracil, which is referred to as CAF; and doxorubicin (Adriamycin) and cyclophosphamide, which is referred to as AC.

[0124] As breast cancer preventative and treatment information can be specifically targeted to subjects in need thereof (*e.g.*, those at risk of developing breast cancer or those that have early signs of breast cancer), provided herein is a method for preventing or reducing the risk of developing breast

cancer in a subject, which comprises: (a) detecting the presence or absence of a polymorphic variation associated with breast cancer at a polymorphic site in a nucleotide sequence in a nucleic acid sample from a subject; (b) identifying a subject with a predisposition to breast cancer, whereby the presence of the polymorphic variation is indicative of a predisposition to breast cancer in the subject; and (c) if such a predisposition is identified, providing the subject with information about methods or products to prevent or reduce breast cancer or to delay the onset of breast cancer. Also provided is a method of targeting information or advertising to a subpopulation of a human population based on the subpopulation being genetically predisposed to a disease or condition, which comprises: (a) detecting the presence or absence of a polymorphic variation associated with breast cancer at a polymorphic site in a nucleotide sequence in a nucleic acid sample from a subject; (b) identifying the subpopulation of subjects in which the polymorphic variation is associated with breast cancer; and (c) providing information only to the subpopulation of subjects about a particular product which may be obtained and consumed or applied by the subject to help prevent or delay onset of the disease or condition.

[0125] Pharmacogenomics methods also may be used to analyze and predict a response to a breast cancer treatment or a drug. For example, if pharmacogenomics analysis indicates a likelihood that an individual will respond positively to a breast cancer treatment with a particular drug, the drug may be administered to the individual. Conversely, if the analysis indicates that an individual is likely to respond negatively to treatment with a particular drug, an alternative course of treatment may be prescribed. A negative response may be defined as either the absence of an efficacious response or the presence of toxic side effects. The response to a therapeutic treatment can be predicted in a background study in which subjects in any of the following populations are genotyped: a population that responds favorably to a treatment regimen, a population that does not respond significantly to a treatment regimen, and a population that responds adversely to a treatment regiment (*e.g.*, exhibits one or more side effects). These populations are provided as examples and other populations and subpopulations may be analyzed. Based upon the results of these analyses, a subject is genotyped to predict whether he or she will respond favorably to a treatment regimen, not respond significantly to a treatment regimen, or respond adversely to a treatment regimen.

[0126] The methods described herein also are applicable to clinical drug trials. One or more polymorphic variants indicative of response to an agent for treating breast cancer or to side effects to an agent for treating breast cancer may be identified using the methods described herein. Thereafter, potential participants in clinical trials of such an agent may be screened to identify those individuals most likely to respond favorably to the drug and exclude those likely to experience side effects. In that way, the effectiveness of drug treatment may be measured in individuals who respond positively to the drug, without lowering the measurement as a result of the inclusion of individuals who are unlikely to respond

positively in the study and without risking undesirable safety problems. In certain embodiments, the agent for treating breast cancer described herein targets *DLG1, KIAA0783, DPF3* or *CENPC1* or a target in the *DLG1, KIAA0783, DPF3* or *CENPC1* pathway.

[0127] Thus, another embodiment is a method of selecting an individual for inclusion in a clinical trial of a treatment or drug comprising the steps of: (a) obtaining a nucleic acid sample from an individual; (b) determining the identity of a polymorphic variation which is associated with a positive response to the treatment or the drug, or at least one polymorphic variation which is associated with a negative response to the treatment or the drug in the nucleic acid sample, and (c) including the individual in the clinical trial if the nucleic acid sample contains said polymorphic variation associated with a positive response to the treatment or the drug or if the nucleic acid sample lacks said polymorphic variation associated with a negative response to the treatment or the drug. In addition, the methods for selecting an individual for inclusion in a clinical trial of a treatment or drug encompass methods with any further limitation described in this disclosure, or those following, specified alone or in any combination. The polymorphic variation may be in a sequence selected individually or in any combination from the group consisting of (i) a polynucleotide sequence set forth in SEQ ID NO: 1-4; (ii) a polynucleotide sequence that is 90% or more identical to a nucleotide sequence set forth in SEQ ID NO: 1-4; (iii) a polynucleotide sequence that encodes a polypeptide having an amino acid sequence identical to or 90% or more identical to an amino acid sequence encoded by a nucleotide sequence set forth in SEQ ID NO: 1-4; and (iv) a fragment of a polynucleotide sequence of (i), (ii), or (iii) comprising the polymorphic site. The including step (c) optionally comprises administering the drug or the treatment to the individual if the nucleic acid sample contains the polymorphic variation associated with a positive response to the treatment or the drug and the nucleic acid sample lacks said biallelic marker associated with a negative response to the treatment or the drug.

[0128] Also provided herein is a method of partnering between a diagnostic/prognostic testing provider and a provider of a consumable product, which comprises: (a) the diagnostic/prognostic testing provider detects the presence or absence of a polymorphic variation associated with breast cancer at a polymorphic site in a nucleotide sequence in a nucleic acid sample from a subject; (b) the diagnostic/prognostic testing provider identifies the subpopulation of subjects in which the polymorphic variation is associated with breast cancer; (c) the diagnostic/prognostic testing provider forwards information to the subpopulation of subjects about a particular product which may be obtained and consumed or applied by the subject to help prevent or delay onset of the disease or condition; and (d) the provider of a consumable product forwards to the diagnostic test provider a fee every time the diagnostic/prognostic test provider forwards information to the subject as set forth in step (c) above.

Compositions Comprising Breast Cancer-Directed Molecules

[0129] Featured herein is a composition comprising a breast cancer cell and one or more molecules specifically directed and targeted to a nucleic acid comprising a *DLG1, KIAA0783, DPF3* or *CENPC1* nucleotide sequence or a *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide. Such directed molecules include, but are not limited to, a compound that binds to a *DLG1, KIAA0783, DPF3* or *CENPC1* nucleic acid or a *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide; a RNAi or siRNA molecule having a strand complementary to a *DLG1, KIAA0783, DPF3* or *CENPC1* nucleotide sequence; an antisense nucleic acid complementary to an RNA encoded by a *DLG1, KIAA0783, DPF3* or *CENPC1* DNA sequence; a ribozyme that hybridizes to a *DLG1, KIAA0783, DPF3* or *CENPC1* nucleotide sequence; a nucleic acid aptamer that specifically binds a *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide; and an antibody that specifically binds to a *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide or binds to a *DLG1, KIAA0783, DPF3* or *CENPC1* nucleic acid. In certain embodiments, the antibody specifically binds to an epitope that comprises a glutamine at amino acid position 278 in SEQ ID NO: 9 or a glycine at amino acid position 389 in SEQ ID NO: 12. In specific embodiments, the breast cancer directed molecule interacts with a *DLG1, KIAA0783, DPF3* or *CENPC1* nucleic acid or polypeptide variant associated with breast cancer. In other embodiments, the breast cancer directed molecule interacts with a polypeptide involved in the *DLG1, KIAA0783, DPF3* or *CENPC1* signal pathway, or a nucleic acid encoding such a polypeptide. Polypeptides involved in the *DLG1, KIAA0783, DPF3* or *CENPC1* signal pathway are discussed herein.

[0130] Compositions sometimes include an adjuvant known to stimulate an immune response, and in certain embodiments, an adjuvant that stimulates a T-cell lymphocyte response. Adjuvants are known, including but not limited to an aluminum adjuvant (*e.g.*, aluminum hydroxide); a cytokine adjuvant or adjuvant that stimulates a cytokine response (*e.g.*, interleukin (IL)-12 and/or γ-interferon cytokines); a Freund-type mineral oil adjuvant emulsion (*e.g.*, Freund's complete or incomplete adjuvant); a synthetic lipoid compound; a copolymer adjuvant (*e.g.*, TitreMax); a saponin; Quil A; a liposome; an oil-in-water emulsion (*e.g.*, an emulsion stabilized by Tween 80 and pluronic polyoxyethlene/polyoxypropylene block copolymer (Syntex Adjuvant Formulation); TitreMax; detoxified endotoxin (MPL) and mycobacterial cell wall components (TDW, CWS) in 2% squalene (Ribi Adjuvant System)); a muramyl dipeptide; an immune-stimulating complex (ISCOM, *e.g.*, an Ag-modified saponin/cholesterol micelle that forms stable cage-like structure); an aqueous phase adjuvant that does not have a depot effect (*e.g.*, Gerbu adjuvant); a carbohydrate polymer (*e.g.*, AdjuPrime); L-tyrosine; a manide-oleate compound (*e.g.*, Montanide); an ethylene-vinyl acetate copolymer (*e.g.*, Elvax 40W1,2); or lipid A, for example. Such compositions are useful for generating an immune response against a breast cancer directed molecule (*e.g.*, an HLA-binding subsequence within a polypeptide encoded by a nucleotide sequence in SEQ ID

NO: 1-4). In such methods, a peptide having an amino acid subsequence of a polypeptide encoded by a nucleotide sequence in SEQ ID NO: 1-4 is delivered to a subject, where the subsequence binds to an HLA molecule and induces a CTL lymphocyte response. The peptide sometimes is delivered to the subject as an isolated peptide or as a minigene in a plasmid that encodes the peptide. Methods for identifying HLA-binding subsequences in such polypeptides are known (*see e.g.*, publication WO02/20616 and PCT application US98/01373 for methods of identifying such sequences).

[0131] The breast cancer cell may be in a group of breast cancer cells and/or other types of cells cultured *in vitro* or in a tissue having breast cancer cells (*e.g.*, a melanocytic lesion) maintained *in vitro* or present in an animal *in vivo* (*e.g.*, a rat, mouse, ape or human). In certain embodiments, a composition comprises a component from a breast cancer cell or from a subject having a breast cancer cell instead of the breast cancer cell or in addition to the breast cancer cell, where the component sometimes is a nucleic acid molecule (*e.g.*, genomic DNA), a protein mixture or isolated protein, for example. The aforementioned compositions have utility in diagnostic, prognostic and pharmacogenomic methods described previously and in breast cancer therapeutics described hereafter. Certain breast cancer molecules are described in greater detail below.

Compounds

[0132] Compounds can be obtained using any of the numerous approaches in combinatorial library methods known in the art, including: biological libraries; peptoid libraries (libraries of molecules having the functionalities of peptides, but with a novel, non-peptide backbone which are resistant to enzymatic degradation but which nevertheless remain bioactive (see, *e.g.*, Zuckermann *et al.*, J. Med. Chem.37: 2678-85 (1994)); spatially addressable parallel solid phase or solution phase libraries; synthetic library methods requiring deconvolution; "one-bead one-compound" library methods; and synthetic library methods using affinity chromatography selection. Biological library and peptoid library approaches are typically limited to peptide libraries, while the other approaches are applicable to peptide, non-peptide oligomer or small molecule libraries of compounds (Lam, Anticancer Drug Des. 12: 145, (1997)). Examples of methods for synthesizing molecular libraries are described, for example, in DeWitt *et al.*, Proc. Natl. Acad. Sci. U.S.A. 90: 6909 (1993); Erb *et al.*, Proc. Natl. Acad. Sci. USA 91: 11422 (1994); Zuckermann *et al.*, J. Med. Chem. 37: 2678 (1994); Cho *et al.*, Science 261: 1303 (1993); Carrell *et al.*, Angew. Chem. Int. Ed. Engl. 33: 2059 (1994); Carell *et al.*, Angew. Chem. Int. Ed. Engl. 33: 2061 (1994); and in Gallop *et al.*, J. Med. Chem. 37: 1233 (1994).

[0133] Libraries of compounds may be presented in solution (*e.g.*, Houghten, Biotechniques 13: 412-421 (1992)), or on beads (Lam, Nature 354: 82-84 (1991)), chips (Fodor, Nature 364: 555-556 (1993)), bacteria or spores (Ladner, United States Patent No. 5,223,409), plasmids (Cull *et al.*, Proc. Natl.

Acad. Sci. USA 89: 1865-1869 (1992)) or on phage (Scott and Smith, Science 249: 386-390 (1990); Devlin, Science 249: 404-406 (1990); Cwirla *et al.*, Proc. Natl. Acad. Sci. 87: 6378-6382 (1990); Felici, J. Mol. Biol. 222: 301-310 (1991); Ladner supra.).

[0134] A compound sometimes alters expression and sometimes alters activity of a *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide and may be a small molecule. Small molecules include, but are not limited to, peptides, peptidomimetics (*e.g.*, peptoids), amino acids, amino acid analogs, polynucleotides, polynucleotide analogs, nucleotides, nucleotide analogs, organic or inorganic compounds (i.e., including heteroorganic and organometallic compounds) having a molecular weight less than about 10,000 grams per mole, organic or inorganic compounds having a molecular weight less than about 5,000 grams per mole, organic or inorganic compounds having a molecular weight less than about 1,000 grams per mole, organic or inorganic compounds having a molecular weight less than about 500 grams per mole, and salts, esters, and other pharmaceutically acceptable forms of such compounds.

Antisense Nucleic Acid Molecules, Ribozymes, RNAi, siRNA and Modified Nucleic Acid Molecules

[0135] An "antisense" nucleic acid refers to a nucleotide sequence complementary to a "sense" nucleic acid encoding a polypeptide, *e.g.*, complementary to the coding strand of a double-stranded cDNA molecule or complementary to an mRNA sequence. The antisense nucleic acid can be complementary to an entire coding strand in SEQ ID NO: 1-8, or to a portion thereof or a substantially identical sequence thereof. In another embodiment, the antisense nucleic acid molecule is antisense to a "noncoding region" of the coding strand of a nucleotide sequence in SEQ ID NO: 1-8 (*e.g.*, 5' and 3' untranslated regions).

[0136] An antisense nucleic acid can be designed such that it is complementary to the entire coding region of an mRNA encoded by a nucleotide sequence in SEQ ID NO: 1-4 (e.g., SEQ ID NO: 6-11), and often the antisense nucleic acid is an oligonucleotide antisense to only a portion of a coding or noncoding region of the mRNA. For example, the antisense oligonucleotide can be complementary to the region surrounding the translation start site of the mRNA, *e.g.*, between the -10 and +10 regions of the target gene nucleotide sequence of interest. An antisense oligonucleotide can be, for example, about 7, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, or more nucleotides in length. The antisense nucleic acids, which include the ribozymes described hereafter, can be designed to target a nucleotide sequence in SEQ ID NO: 1-8, often a variant associated with breast cancer, or a substantially identical sequence thereof. Among the variants, minor alleles and major alleles can be targeted, and those associated with a higher risk of breast cancer are often designed, tested, and administered to subjects.

[0137] An antisense nucleic acid can be constructed using chemical synthesis and enzymatic ligation reactions using standard procedures. For example, an antisense nucleic acid (*e.g.*, an antisense oligonucleotide) can be chemically synthesized using naturally occurring nucleotides or variously modified nucleotides designed to increase the biological stability of the molecules or to increase the physical stability of the duplex formed between the antisense and sense nucleic acids, *e.g.*, phosphorothioate derivatives and acridine substituted nucleotides can be used. Antisense nucleic acid also can be produced biologically using an expression vector into which a nucleic acid has been subcloned in an antisense orientation (i.e., RNA transcribed from the inserted nucleic acid will be of an antisense orientation to a target nucleic acid of interest, described further in the following subsection).

[0138] When utilized as therapeutics, antisense nucleic acids typically are administered to a subject (*e.g.*, by direct injection at a tissue site) or generated in situ such that they hybridize with or bind to cellular mRNA and/or genomic DNA encoding a polypeptide and thereby inhibit expression of the polypeptide, for example, by inhibiting transcription and/or translation. Alternatively, antisense nucleic acid molecules can be modified to target selected cells and then are administered systemically. For systemic administration, antisense molecules can be modified such that they specifically bind to receptors or antigens expressed on a selected cell surface, for example, by linking antisense nucleic acid molecules to peptides or antibodies which bind to cell surface receptors or antigens. Antisense nucleic acid molecules can also be delivered to cells using the vectors described herein. Sufficient intracellular concentrations of antisense molecules are achieved by incorporating a strong promoter, such as a pol II or pol III promoter, in the vector construct.

[0139] Antisense nucleic acid molecules sometimes are *-anomeric nucleic acid molecules. An *-anomeric nucleic acid molecule forms specific double-stranded hybrids with complementary RNA in which, contrary to the usual *-units, the strands run parallel to each other (Gaultier *et al.*, Nucleic Acids. Res. 15: 6625-6641 (1987)). Antisense nucleic acid molecules can also comprise a 2'-o-methylribonucleotide (Inoue *et al.*, Nucleic Acids Res. 15: 6131-6148 (1987)) or a chimeric RNA-DNA analogue (Inoue *et al.*, FEBS Lett. 215: 327-330 (1987)). Antisense nucleic acids sometimes are composed of DNA or PNA or any other nucleic acid derivatives described previously.

[0140] In another embodiment, an antisense nucleic acid is a ribozyme. A ribozyme having specificity for a *DLG1, KIAA0783, DPF3* or *CENPC1* nucleotide sequence can include one or more sequences complementary to such a nucleotide sequence, and a sequence having a known catalytic region responsible for mRNA cleavage (see *e.g.*, U.S. Pat. No. 5,093,246 or Haselhoff and Gerlach, Nature 334: 585-591 (1988)). For example, a derivative of a Tetrahymena L-19 IVS RNA is sometimes utilized in which the nucleotide sequence of the active site is complementary to the nucleotide sequence to be cleaved in a mRNA (see *e.g.*, Cech *et al.* U.S. Patent No. 4,987,071; and Cech *et al.* U.S. Patent No.

5,116,742). Also, target mRNA sequences can be used to select a catalytic RNA having a specific ribonuclease activity from a pool of RNA molecules (see *e.g.*, Bartel & Szostak, Science 261: 1411-1418 (1993)).

[0141] Breast cancer directed molecules include in certain embodiments nucleic acids that can form triple helix structures with a *DLG1, KIAA0783, DPF3* or *CENPC1* nucleotide sequence or a substantially identical sequence thereof, especially one that includes a regulatory region that controls expression of a polypeptide. Gene expression can be inhibited by targeting nucleotide sequences complementary to the regulatory region of a *DLG1, KIAA0783, DPF3* or *CENPC1* nucleotide sequence or a substantially identical sequence (*e.g.*, promoter and/or enhancers) to form triple helical structures that prevent transcription of a gene in target cells (see *e.g.*, Helene, Anticancer Drug Des. 6(6): 569-84 (1991); Helene *et al.*, Ann. N.Y. Acad. Sci. 660: 27-36 (1992); and Maher, Bioassays 14(12): 807-15 (1992). Potential sequences that can be targeted for triple helix formation can be increased by creating a so-called "switchback" nucleic acid molecule. Switchback molecules are synthesized in an alternating 5'-3', 3'-5' manner, such that they base pair with first one strand of a duplex and then the other, eliminating the necessity for a sizeable stretch of either purines or pyrimidines to be present on one strand of a duplex.

[0142] Breast cancer directed molecules include RNAi and siRNA nucleic acids. Gene expression may be inhibited by the introduction of double-stranded RNA (dsRNA), which induces potent and specific gene silencing, a phenomenon called RNA interference or RNAi. See, *e.g.*, Fire *et al.*, US Patent Number 6,506,559; Tuschl *et al.* PCT International Publication No. WO 01/75164; Kay *et al.* PCT International Publication No. WO 03/010180A1; or Bosher JM, Labouesse, Nat Cell Biol 2000 Feb;2(2):E31-6. This process has been improved by decreasing the size of the double-stranded RNA to 20-24 base pairs (to create small-interfering RNAs or siRNAs) that "switched off" genes in mammalian cells without initiating an acute phase response, i.e., a host defense mechanism that often results in cell death (see, *e.g.*, Caplen *et al.* Proc Natl Acad Sci U S A. 2001 Aug 14;98(17):9742-7 and Elbashir *et al.* Methods 2002 Feb;26(2):199-213). There is increasing evidence of post-transcriptional gene silencing by RNA interference (RNAi) for inhibiting targeted expression in mammalian cells at the mRNA level, in human cells. There is additional evidence of effective methods for inhibiting the proliferation and migration of tumor cells in human patients, and for inhibiting metastatic cancer development (see, *e.g.*, U.S. Patent Application No. US2001000993183; Caplen *et al.* Proc Natl Acad Sci U S A; and Abderrahmani *et al.* Mol Cell Biol 2001 Nov21(21):7256-67).

[0143] An "siRNA" or "RNAi" refers to a nucleic acid that forms a double stranded RNA and has the ability to reduce or inhibit expression of a gene or target gene when the siRNA is delivered to or expressed in the same cell as the gene or target gene. "siRNA" refers to short double-stranded RNA formed by the complementary strands. Complementary portions of the siRNA that hybridize to form the

double stranded molecule often have substantial or complete identity to the target molecule sequence. In one embodiment, an siRNA refers to a nucleic acid that has substantial or complete identity to a target gene and forms a double stranded siRNA.

[0144] When designing the siRNA molecules, the targeted region often is selected from a given DNA sequence beginning 50 to 100 nucleotides downstream of the start codon. See, *e.g.*, Elbashir et al,. Methods 26:199-213 (2002). Initially, 5' or 3' UTRs and regions nearby the start codon were avoided assuming that UTR-binding proteins and/or translation initiation complexes may interfere with binding of the siRNP or RISC endonuclease complex. Sometimes regions of the target 23 nucleotides in length conforming to the sequence motif AA(N19)TT (N, an nucleotide), and regions with approximately 30% to 70% G/C-content (often about 50% G/C-content) often are selected. If no suitable sequences are found, the search often is extended using the motif NA(N21). The sequence of the sense siRNA sometimes corresponds to (N19) TT or N21 (position 3 to 23 of the 23-nt motif), respectively. In the latter case, the 3' end of the sense siRNA often is converted to TT. The rationale for this sequence conversion is to generate a symmetric duplex with respect to the sequence composition of the sense and antisense 3' overhangs. The antisense siRNA is synthesized as the complement to position 1 to 21 of the 23-nt motif. Because position 1 of the 23-nt motif is not recognized sequence-specifically by the antisense siRNA, the 3'-most nucleotide residue of the antisense siRNA can be chosen deliberately. However, the penultimate nucleotide of the antisense siRNA (complementary to position 2 of the 23-nt motif) often is complementary to the targeted sequence. For simplifying chemical synthesis, TT often is utilized. siRNAs corresponding to the target motif NAR(N17)YNN, where R is purine (A,G) and Y is pyrimidine (C,U), often are selected. Respective 21 nucleotide sense and antisense siRNAs often begin with a purine nucleotide and can also be expressed from pol III expression vectors without a change in targeting site. Expression of RNAs from pol III promoters often is efficient when the first transcribed nucleotide is a purine.

[0145] The sequence of the siRNA can correspond to the full length target gene, or a subsequence thereof. Often, the siRNA is about 15 to about 50 nucleotides in length (*e.g.*, each complementary sequence of the double stranded siRNA is 15-50 nucleotides in length, and the double stranded siRNA is about 15-50 base pairs in length, sometimes about 20-30 nucleotides in length or about 20-25 nucleotides in length, *e.g.*, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, or 30 nucleotides in length. The siRNA sometimes is about 21 nucleotides in length. Methods of using siRNA are well known in the art, and specific siRNA molecules may be purchased from a number of companies including Dharmacon Research, Inc.

[0146] Antisense, ribozyme, RNAi and siRNA nucleic acids can be altered to form modified nucleic acid molecules. The nucleic acids can be altered at base moieties, sugar moieties or phosphate backbone moieties to improve stability, hybridization, or solubility of the molecule. For example, the deoxyribose

phosphate backbone of nucleic acid molecules can be modified to generate peptide nucleic acids (see Hyrup *et al.*, Bioorganic & Medicinal Chemistry 4 (1): 5-23 (1996)). As used herein, the terms "peptide nucleic acid" or "PNA" refers to a nucleic acid mimic such as a DNA mimic, in which the deoxyribose phosphate backbone is replaced by a pseudopeptide backbone and only the four natural nucleobases are retained. The neutral backbone of a PNA can allow for specific hybridization to DNA and RNA under conditions of low ionic strength. Synthesis of PNA oligomers can be performed using standard solid phase peptide synthesis protocols as described, for example, in Hyrup *et al.*, (1996) supra and Perry-O'Keefe *et al.*, Proc. Natl. Acad. Sci. 93: 14670-675 (1996).

[0147] PNA nucleic acids can be used in prognostic, diagnostic, and therapeutic applications. For example, PNAs can be used as antisense or antigene agents for sequence-specific modulation of gene expression by, for example, inducing transcription or translation arrest or inhibiting replication. PNA nucleic acid molecules can also be used in the analysis of single base pair mutations in a gene, (*e.g.*, by PNA-directed PCR clamping); as "artificial restriction enzymes" when used in combination with other enzymes, (*e.g.*, S1 nucleases (Hyrup (1996) supra)); or as probes or primers for DNA sequencing or hybridization (Hyrup *et al.*, (1996) supra; Perry-O'Keefe supra).

[0148] In other embodiments, oligonucleotides may include other appended groups such as peptides (*e.g.*, for targeting host cell receptors *in vivo*), or agents facilitating transport across cell membranes (see *e.g.*, Letsinger *et al.*, Proc. Natl. Acad. Sci. USA 86: 6553-6556 (1989); Lemaitre *et al.*, Proc. Natl. Acad. Sci. USA 84: 648-652 (1987); PCT Publication No. W088/09810) or the blood-brain barrier (see, *e.g.*, PCT Publication No. W089/10134). In addition, oligonucleotides can be modified with hybridization-triggered cleavage agents (See, *e.g.*, Krol *et al.*, Bio-Techniques 6: 958-976 (1988)) or intercalating agents. (See, *e.g.*, Zon, Pharm. Res. 5: 539-549 (1988)). To this end, the oligonucleotide may be conjugated to another molecule, (*e.g.*, a peptide, hybridization triggered cross-linking agent, transport agent, or hybridization-triggered cleavage agent).

[0149] Also included herein are molecular beacon oligonucleotide primer and probe molecules having one or more regions complementary to a nucleotide sequence of SEQ ID NO: 1-8 or a substantially identical sequence thereof, two complementary regions one having a fluorophore and one a quencher such that the molecular beacon is useful for quantifying the presence of the nucleic acid in a sample. Molecular beacon nucleic acids are described, for example, in Lizardi *et al.*, U.S. Patent No. 5,854,033; Nazarenko *et al.*, U.S. Patent No. 5,866,336, and Livak *et al.*, U.S. Patent 5,876,930.

Antibodies

[0150] The term "antibody" as used herein refers to an immunoglobulin molecule or immunologically active portion thereof, i.e., an antigen-binding portion. Examples of immunologically

active portions of immunoglobulin molecules include F(ab) and F(ab')$_2$ fragments which can be generated by treating the antibody with an enzyme such as pepsin. An antibody sometimes is a polyclonal, monoclonal, recombinant (*e.g.*, a chimeric or humanized), fully human, non-human (*e.g.*, murine), or a single chain antibody. An antibody may have effector function and can fix complement, and is sometimes coupled to a toxin or imaging agent.

[0151] A full-length polypeptide or antigenic peptide fragment encoded by a *DLG1, KIAA0783, DPF3* or *CENPC1* nucleotide sequence can be used as an immunogen or can be used to identify antibodies made with other immunogens, *e.g.*, cells, membrane preparations, and the like. An antigenic peptide often includes at least 8 amino acid residues of the amino acid sequences encoded by a nucleotide sequence of SEQ ID NO: 1-8, or substantially identical sequence thereof, and encompasses an epitope. Antigenic peptides sometimes include 10 or more amino acids, 15 or more amino acids, 20 or more amino acids, or 30 or more amino acids. Hydrophilic and hydrophobic fragments of polypeptides sometimes are used as immunogens.

[0152] Epitopes encompassed by the antigenic peptide are regions located on the surface of the polypeptide (*e.g.*, hydrophilic regions) as well as regions with high antigenicity. For example, an Emini surface probability analysis of the human polypeptide sequence can be used to indicate the regions that have a particularly high probability of being localized to the surface of the polypeptide and are thus likely to constitute surface residues useful for targeting antibody production. The antibody may bind an epitope on any domain or region on polypeptides described herein.

[0153] Also, chimeric, humanized, and completely human antibodies are useful for applications which include repeated administration to subjects. Chimeric and humanized monoclonal antibodies, comprising both human and non-human portions, can be made using standard recombinant DNA techniques. Such chimeric and humanized monoclonal antibodies can be produced by recombinant DNA techniques known in the art, for example using methods described in Robinson et al International Application No. PCT/US86/02269; Akira, et al European Patent Application 184,187; Taniguchi, M., European Patent Application 171,496; Morrison et al European Patent Application 173,494; Neuberger et al PCT International Publication No. WO 86/01533; Cabilly et al U.S. Patent No. 4,816,567; Cabilly et al European Patent Application 125,023; Better *et al.*, Science 240: 1041-1043 (1988); Liu *et al.*, Proc. Natl. Acad. Sci. USA 84: 3439-3443 (1987); Liu *et al.*, J. Immunol. 139: 3521-3526 (1987); Sun *et al.*, Proc. Natl. Acad. Sci. USA 84: 214-218 (1987); Nishimura *et al.*, Canc. Res. 47: 999-1005 (1987); Wood *et al.*, Nature 314: 446-449 (1985); and Shaw *et al.*, J. Natl. Cancer Inst. 80: 1553-1559 (1988); Morrison, S. L., Science 229: 1202-1207 (1985); Oi *et al.*, BioTechniques 4: 214 (1986); Winter U.S. Patent 5,225,539; Jones *et al.*, Nature 321: 552-525 (1986); Verhoeyan *et al.*, Science 239: 1534; and Beidler *et al.*, J. Immunol. 141: 4053-4060 (1988).

[0154] Completely human antibodies are particularly desirable for therapeutic treatment of human patients. Such antibodies can be produced using transgenic mice that are incapable of expressing endogenous immunoglobulin heavy and light chains genes, but which can express human heavy and light chain genes. See, for example, Lonberg and Huszar, Int. Rev. Immunol. 13: 65-93 (1995); and U.S. Patent Nos. 5,625,126; 5,633,425; 5,569,825; 5,661,016; and 5,545,806. In addition, companies such as Abgenix, Inc. (Fremont, CA) and Medarex, Inc. (Princeton, NJ), can be engaged to provide human antibodies directed against a selected antigen using technology similar to that described above. Completely human antibodies that recognize a selected epitope also can be generated using a technique referred to as "guided selection." In this approach a selected non-human monoclonal antibody (*e.g.*, a murine antibody) is used to guide the selection of a completely human antibody recognizing the same epitope. This technology is described for example by Jespers *et al.*, Bio/Technology 12: 899-903 (1994).

[0155] Antibody can be a single chain antibody. A single chain antibody (scFV) can be engineered (see, *e.g.*, Colcher *et al.*, Ann. N Y Acad. Sci. 880: 263-80 (1999); and Reiter, Clin. Cancer Res. 2: 245-52 (1996)). Single chain antibodies can be dimerized or multimerized to generate multivalent antibodies having specificities for different epitopes of the same target polypeptide.

[0156] Antibodies also may be selected or modified so that they exhibit reduced or no ability to bind an Fc receptor. For example, an antibody may be an isotype or subtype, fragment or other mutant, which does not support binding to an Fc receptor (*e.g.*, it has a mutagenized or deleted Fc receptor binding region).

[0157] Also, an antibody (or fragment thereof) may be conjugated to a therapeutic moiety such as a cytotoxin, a therapeutic agent or a radioactive metal ion. A cytotoxin or cytotoxic agent includes any agent that is detrimental to cells. Examples include taxol, cytochalasin B, gramicidin D, ethidium bromide, emetine, mitomycin, etoposide, tenoposide, vincristine, vinblastine, colchicin, doxorubicin, daunorubicin, dihydroxy anthracin dione, mitoxantrone, mithramycin, actinomycin D, 1 dehydrotestosterone, glucocorticoids, procaine, tetracaine, lidocaine, propranolol, and puromycin and analogs or homologs thereof. Therapeutic agents include, but are not limited to, antimetabolites (*e.g.*, methotrexate, 6-mercaptopurine, 6-thioguanine, cytarabine, 5-fluorouracil decarbazine), alkylating agents (*e.g.*, mechlorethamine, thiotepa chlorambucil, melphalan, carmustine (BCNU) and lomustine (CCNU), cyclophosphamide, busulfan, dibromomannitol, streptozotocin, mitomycin C, and cis-dichlorodiamine platinum (II) (DDP) cisplatin), anthracyclines (*e.g.*, daunorubicin (formerly daunomycin) and doxorubicin), antibiotics (*e.g.*, dactinomycin (formerly actinomycin), bleomycin, mithramycin, and anthramycin (AMC)), and anti-mitotic agents (*e.g.*, vincristine and vinblastine).

[0158] Antibody conjugates can be used for modifying a given biological response. For example, the drug moiety may be a protein or polypeptide possessing a desired biological activity. Such proteins

may include, for example, a toxin such as abrin, ricin A, pseudomonas exotoxin, or diphtheria toxin; a polypeptide such as tumor necrosis factor, γ-interferon, α-interferon, nerve growth factor, platelet derived growth factor, tissue plasminogen activator; or, biological response modifiers such as, for example, lymphokines, interleukin-1 ("IL-1"), interleukin-2 ("IL-2"), interleukin-6 ("IL-6"), granulocyte macrophage colony stimulating factor ("GM-CSF"), granulocyte colony stimulating factor ("G-CSF"), or other growth factors. Also, an antibody can be conjugated to a second antibody to form an antibody heteroconjugate as described by Segal in U.S. Patent No. 4,676,980, for example.

[0159] An antibody (*e.g.*, monoclonal antibody) can be used to isolate target polypeptides by standard techniques, such as affinity chromatography or immunoprecipitation. Moreover, an antibody can be used to detect a target polypeptide (*e.g.*, in a cellular lysate or cell supernatant) in order to evaluate the abundance and pattern of expression of the polypeptide. Antibodies can be used diagnostically to monitor polypeptide levels in tissue as part of a clinical testing procedure, *e.g.*, to determine the efficacy of a given treatment regimen. Detection can be facilitated by coupling (i.e., physically linking) the antibody to a detectable substance (i.e., antibody labeling). Examples of detectable substances include various enzymes, prosthetic groups, fluorescent materials, luminescent materials, bioluminescent materials, and radioactive materials. Examples of suitable enzymes include horseradish peroxidase, alkaline phosphatase, β-galactosidase, or acetylcholinesterase; examples of suitable prosthetic group complexes include streptavidin/biotin and avidin/biotin; examples of suitable fluorescent materials include umbelliferone, fluorescein, fluorescein isothiocyanate, rhodamine, dichlorotriazinylamine fluorescein, dansyl chloride or phycoerythrin; an example of a luminescent material includes luminol; examples of bioluminescent materials include luciferase, luciferin, and aequorin, and examples of suitable radioactive material include ^{125}I, ^{131}I, ^{35}S or ^{3}H. Also, an antibody can be utilized as a test molecule for determining whether it can treat breast cancer, and as a therapeutic for administration to a subject for treating breast cancer.

[0160] An antibody can be made by immunizing with a purified antigen, or a fragment thereof, *e.g.*, a fragment described herein, a membrane associated antigen, tissues, *e.g.*, crude tissue preparations, whole cells, preferably living cells, lysed cells, or cell fractions.

[0161] Included herein are antibodies which bind only a native polypeptide, only denatured or otherwise non-native polypeptide, or which bind both, as well as those having linear or conformational epitopes. Conformational epitopes sometimes can be identified by selecting antibodies that bind to native but not denatured polypeptide. Also featured are antibodies that specifically bind to a polypeptide variant associated with breast cancer.

Screening Assays

[0162] Featured herein are methods for identifying a candidate therapeutic for treating breast cancer. The methods comprise contacting a test molecule with a target molecule in a system. A "target molecule" as used herein refers to a nucleic acid of SEQ ID NO: 1-8, a substantially identical nucleic acid thereof, or a fragment thereof, and an encoded polypeptide of the foregoing. The method also comprises determining the presence or absence of an interaction between the test molecule and the target molecule, where the presence of an interaction between the test molecule and the nucleic acid or polypeptide identifies the test molecule as a candidate breast cancer therapeutic. The interaction between the test molecule and the target molecule may be quantified.

[0163] Test molecules and candidate therapeutics include, but are not limited to, compounds, antisense nucleic acids, siRNA molecules, ribozymes, polypeptides or proteins encoded by a *DLG1, KIAA0783, DPF3* or *CENPC1* nucleic acids, or a substantially identical sequence or fragment thereof, and immunotherapeutics (*e.g.*, antibodies and HLA-presented polypeptide fragments). A test molecule or candidate therapeutic may act as a modulator of target molecule concentration or target molecule function in a system. A "modulator" may agonize (i.e., up-regulates) or antagonize (i.e., down-regulates) a target molecule concentration partially or completely in a system by affecting such cellular functions as DNA replication and/or DNA processing (*e.g.*, DNA methylation or DNA repair), RNA transcription and/or RNA processing (*e.g.*, removal of intronic sequences and/or translocation of spliced mRNA from the nucleus), polypeptide production (*e.g.*, translation of the polypeptide from mRNA), and/or polypeptide post-translational modification (*e.g.*, glycosylation, phosphorylation, and proteolysis of pro-polypeptides). A modulator may also agonize or antagonize a biological function of a target molecule partially or completely, where the function may include adopting a certain structural conformation, interacting with one or more binding partners, ligand binding, catalysis (*e.g.*, phosphorylation, dephosphorylation, hydrolysis, methylation, and isomerization), and an effect upon a cellular event (*e.g.*, effecting progression of breast cancer).

[0164] As used herein, the term "system" refers to a cell free *in vitro* environment and a cell-based environment such as a collection of cells, a tissue, an organ, or an organism. A system is "contacted" with a test molecule in a variety of manners, including adding molecules in solution and allowing them to interact with one another by diffusion, cell injection, and any administration routes in an animal. As used herein, the term "interaction" refers to an effect of a test molecule on test molecule, where the effect sometimes is binding between the test molecule and the target molecule, and sometimes is an observable change in cells, tissue, or organism.

[0165] There are many standard methods for detecting the presence or absence of an interaction between a test molecule and a target molecule. For example, titrametric, acidimetric, radiometric, NMR,

monolayer, polarographic, spectrophotometric, fluorescent, and ESR assays probative of a target molecule interaction may be utilized.

[0166] In general, an interaction can be determined by labeling the test molecule and/or the *DLG1, KIAA0783, DPF3* or *CENPC1* molecule, where the label is covalently or non-covalently attached to the test molecule or *DLG1, KIAA0783, DPF3* or *CENPC1* molecule. The label is sometimes a radioactive molecule such as ^{125}I, ^{131}I, ^{35}S or ^{3}H, which can be detected by direct counting of radioemission or by scintillation counting. Also, enzymatic labels such as horseradish peroxidase, alkaline phosphatase, or luciferase may be utilized where the enzymatic label can be detected by determining conversion of an appropriate substrate to product. Also, presence or absence of an interaction can be determined without labeling. For example, a microphysiometer (e.g., Cytosensor) is an analytical instrument that measures the rate at which a cell acidifies its environment using a light-addressable potentiometric sensor (LAPS). Changes in this acidification rate can be used as an indication of an interaction between a test molecule and *DLG1, KIAA0783, DPF3* or *CENPC1* (McConnell, H. M. et al., Science 257: 1906-1912 (1992)).

[0167] In cell-based systems, cells typically include a *DLG1, KIAA0783, DPF3* or *CENPC1* nucleic acid or polypeptide or variants thereof and are often of mammalian origin, although the cell can be of any origin. Whole cells, cell homogenates, and cell fractions (e.g., cell membrane fractions) can be subjected to analysis. Where interactions between a test molecule with a *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide or variant thereof are monitored, soluble and/or membrane bound forms of the polypeptide or variant may be utilized. Where membrane-bound forms of the polypeptide are used, it may be desirable to utilize a solubilizing agent. Examples of such solubilizing agents include non-ionic detergents such as n-octylglucoside, n-dodecylglucoside, n-dodecylmaltoside, octanoyl-N-methylglucamide, decanoyl-N-methylglucamide, Triton® X-100, Triton® X-114, Thesit®, Isotridecypoly(ethylene glycol ether)n, 3-[(3-cholamidopropyl)dimethylamminio]-1-propane sulfonate (CHAPS), 3-[(3-cholamidopropyl)dimethylamminio]-2-hydroxy-1-propane sulfonate (CHAPSO), or N-dodecyl-N,N-dimethyl-3-ammonio-1-propane sulfonate.

[0168] An interaction between two molecules also can be detected by monitoring fluorescence energy transfer (FET) (see, for example, Lakowicz et al., U.S. Patent No. 5,631,169; Stavrianopoulos et al. U.S. Patent No. 4,868,103). A fluorophore label on a first, "donor" molecule is selected such that its emitted fluorescent energy will be absorbed by a fluorescent label on a second, "acceptor" molecule, which in turn is able to fluoresce due to the absorbed energy. Alternately, the "donor" polypeptide molecule may simply utilize the natural fluorescent energy of tryptophan residues. Labels are chosen that emit different wavelengths of light, such that the "acceptor" molecule label may be differentiated from that of the "donor". Since the efficiency of energy transfer between the labels is related to the distance separating the molecules, the spatial relationship between the molecules can be assessed. In a

situation in which binding occurs between the molecules, the fluorescent emission of the "acceptor" molecule label in the assay should be maximal. An FET binding event can be conveniently measured through standard fluorometric detection means well known in the art (e.g., using a fluorimeter).

[0169] In another embodiment, determining the presence or absence of an interaction between a test molecule and a *DLG1, KIAA0783, DPF3* or *CENPC1* molecule can be effected by using real-time Biomolecular Interaction Analysis (BIA) (see, e.g., Sjolander & Urbaniczk, Anal. Chem. 63: 2338-2345 (1991) and Szabo et al., Curr. Opin. Struct. Biol. 5: 699-705 (1995)). "Surface plasmon resonance" or "BIA" detects biospecific interactions in real time, without labeling any of the interactants (e.g., BIAcore). Changes in the mass at the binding surface (indicative of a binding event) result in alterations of the refractive index of light near the surface (the optical phenomenon of surface plasmon resonance (SPR)), resulting in a detectable signal which can be used as an indication of real-time reactions between biological molecules.

[0170] In another embodiment, the *DLG1, KIAA0783, DPF3* or *CENPC1* molecule or test molecules are anchored to a solid phase. The *DLG1, KIAA0783, DPF3* or *CENPC1* molecule/test molecule complexes anchored to the solid phase can be detected at the end of the reaction. The target *DLG1, KIAA0783, DPF3* or *CENPC1* molecule is often anchored to a solid surface, and the test molecule, which is not anchored, can be labeled, either directly or indirectly, with detectable labels discussed herein.

[0171] It may be desirable to immobilize a *DLG1, KIAA0783, DPF3* or *CENPC1* molecule, an anti-*DLG1, KIAA0783, DPF3* or *CENPC1* antibody, or test molecules to facilitate separation of complexed from uncomplexed forms of *DLG1, KIAA0783, DPF3* or *CENPC1* molecules and test molecules, as well as to accommodate automation of the assay. Binding of a test molecule to a *DLG1, KIAA0783, DPF3* or *CENPC1* molecule can be accomplished in any vessel suitable for containing the reactants. Examples of such vessels include microtiter plates, test tubes, and micro-centrifuge tubes. In one embodiment, a fusion polypeptide can be provided which adds a domain that allows a *DLG1, KIAA0783, DPF3* or *CENPC1* molecule to be bound to a matrix. For example, glutathione-S-transferase/*DLG1, KIAA0783, DPF3* or *CENPC1* fusion polypeptides or glutathione-S-transferase/target fusion polypeptides can be adsorbed onto glutathione sepharose beads (Sigma Chemical, St. Louis, MO) or glutathione derivitized microtiter plates, which are then combined with the test compound or the test compound and either the non-adsorbed target polypeptide or *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide, and the mixture incubated under conditions conducive to complex formation (e.g., at physiological conditions for salt and pH). Following incubation, the beads or microtiter plate wells are washed to remove any unbound components, the matrix immobilized in the case of beads, complex determined either directly or indirectly, for example, as described above. Alternatively, the complexes can be dissociated from the

matrix, and the level of *DLG1, KIAA0783, DPF3* or *CENPC1* binding or activity determined using standard techniques.

[0172] Other techniques for immobilizing a *DLG1, KIAA0783, DPF3* or *CENPC1* molecule on matrices include using biotin and streptavidin. For example, biotinylated *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide or target molecules can be prepared from biotin-NHS (N-hydroxy-succinimide) using techniques known in the art (e.g., biotinylation kit, Pierce Chemicals, Rockford, IL), and immobilized in the wells of streptavidin-coated 96 well plates (Pierce Chemical).

[0173] In order to conduct the assay, the non-immobilized component is added to the coated surface containing the anchored component. After the reaction is complete, unreacted components are removed (e.g., by washing) under conditions such that any complexes formed will remain immobilized on the solid surface. The detection of complexes anchored on the solid surface can be accomplished in a number of ways. Where the previously non-immobilized component is pre-labeled, the detection of label immobilized on the surface indicates that complexes were formed. Where the previously non-immobilized component is not pre-labeled, an indirect label can be used to detect complexes anchored on the surface; e.g., using a labeled antibody specific for the immobilized component (the antibody, in turn, can be directly labeled or indirectly labeled with, e.g., a labeled anti-Ig antibody).

[0174] In one embodiment, this assay is performed utilizing antibodies reactive with *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide or test molecules but which do not interfere with binding of the *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide to its test molecule. Such antibodies can be derivitized to the wells of the plate, and unbound target or *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide trapped in the wells by antibody conjugation. Methods for detecting such complexes, in addition to those described above for the GST-immobilized complexes, include immunodetection of complexes using antibodies reactive with the *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide or target molecule, as well as enzyme-linked assays which rely on detecting an enzymatic activity associated with the *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide or test molecule.

[0175] Alternatively, cell free assays can be conducted in a liquid phase. In such an assay, the reaction products are separated from unreacted components, by any of a number of standard techniques, including but not limited to: differential centrifugation (see, for example, Rivas, G., and Minton, A. P., Trends Biochem Sci Aug;18(8): 284-7 (1993)); chromatography (gel filtration chromatography, ion-exchange chromatography); electrophoresis (see, e.g., Ausubel et al., eds. Current Protocols in Molecular Biology , J. Wiley: New York (1999)); and immunoprecipitation (see, for example, Ausubel, F. et al., eds. Current Protocols in Molecular Biology , J. Wiley: New York (1999)). Such resins and chromatographic techniques are known to one skilled in the art (see, e.g., Heegaard, J Mol. Recognit. Winter; 11(1-6): 141-8 (1998); Hage & Tweed, J. Chromatogr. B Biomed. Sci. Appl. Oct 10; 699 (1-2):

499-525 (1997)). Further, fluorescence energy transfer may also be conveniently utilized, as described herein, to detect binding without further purification of the complex from solution.

[0176] In another embodiment, modulators of *DLG1, KIAA0783, DPF3* or *CENPC1* expression are identified. For example, a cell or cell free mixture is contacted with a candidate compound and the expression of *DLG1, KIAA0783, DPF3* or *CENPC1* mRNA or polypeptide evaluated relative to the level of expression of *DLG1, KIAA0783, DPF3* or *CENPC1* mRNA or polypeptide in the absence of the candidate compound. When expression of *DLG1, KIAA0783, DPF3* or *CENPC1* mRNA or polypeptide is greater in the presence of the candidate compound than in its absence, the candidate compound is identified as a stimulator of *DLG1, KIAA0783, DPF3* or *CENPC1* mRNA or polypeptide expression. Alternatively, when expression of *DLG1, KIAA0783, DPF3* or *CENPC1* mRNA or polypeptide is less (statistically significantly less) in the presence of the candidate compound than in its absence, the candidate compound is identified as an inhibitor of *DLG1, KIAA0783, DPF3* or *CENPC1* mRNA or polypeptide expression. The level of *DLG1, KIAA0783, DPF3* or *CENPC1* mRNA or polypeptide expression can be determined by methods described herein for detecting *DLG1, KIAA0783, DPF3* or *CENPC1* mRNA or polypeptide.

[0177] In another embodiment, binding partners that interact with a *DLG1, KIAA0783, DPF3* or *CENPC1* molecule are detected. The *DLG1, KIAA0783, DPF3* or *CENPC1* molecules can interact with one or more cellular or extracellular macromolecules, such as polypeptides, in vivo, and these molecules that interact with *DLG1, KIAA0783, DPF3* or *CENPC1* molecules are referred to herein as "binding partners." Molecules that disrupt such interactions can be useful in regulating the activity of the target gene product. Such molecules can include, but are not limited to molecules such as antibodies, peptides, and small molecules. Target genes/products for use in this embodiment often are the *DLG1, KIAA0783, DPF3* or *CENPC1* genes herein identified. In an alternative embodiment, provided is a method for determining the ability of the test compound to modulate the activity of a *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide through modulation of the activity of a downstream effector of a *DLG1, KIAA0783, DPF3* or *CENPC1* target molecule. For example, the activity of the effector molecule on an appropriate target can be determined, or the binding of the effector to an appropriate target can be determined, as previously described.

[0178] To identify compounds that interfere with the interaction between the target gene product and its cellular or extracellular binding partner(s), e.g., a substrate, a reaction mixture containing the target gene product and the binding partner is prepared, under conditions and for a time sufficient, to allow the two products to form complex. In order to test an inhibitory agent, the reaction mixture is provided in the presence and absence of the test compound. The test compound can be initially included in the reaction mixture, or can be added at a time subsequent to the addition of the target gene and its

cellular or extracellular binding partner. Control reaction mixtures are incubated without the test compound or with a placebo. The formation of any complexes between the target gene product and the cellular or extracellular binding partner is then detected. The formation of a complex in the control reaction, but not in the reaction mixture containing the test compound, indicates that the compound interferes with the interaction of the target gene product and the interactive binding partner. Additionally, complex formation within reaction mixtures containing the test compound and normal target gene product can also be compared to complex formation within reaction mixtures containing the test compound and mutant target gene product. This comparison can be important in those cases where it is desirable to identify compounds that disrupt interactions of mutant but not normal target gene products.

[0179] These assays can be conducted in a heterogeneous or homogeneous format. Heterogeneous assays involve anchoring either the target gene product or the binding partner onto a solid phase, and detecting complexes anchored on the solid phase at the end of the reaction. In homogeneous assays, the entire reaction is carried out in a liquid phase. In either approach, the order of addition of reactants can be varied to obtain different information about the compounds being tested. For example, test compounds that interfere with the interaction between the target gene products and the binding partners, e.g., by competition, can be identified by conducting the reaction in the presence of the test substance. Alternatively, test compounds that disrupt preformed complexes, e.g., compounds with higher binding constants that displace one of the components from the complex, can be tested by adding the test compound to the reaction mixture after complexes have been formed. The various formats are briefly described below.

[0180] In a heterogeneous assay system, either the target gene product or the interactive cellular or extracellular binding partner, is anchored onto a solid surface (e.g., a microtiter plate), while the non-anchored species is labeled, either directly or indirectly. The anchored species can be immobilized by non-covalent or covalent attachments. Alternatively, an immobilized antibody specific for the species to be anchored can be used to anchor the species to the solid surface.

[0181] In order to conduct the assay, the partner of the immobilized species is exposed to the coated surface with or without the test compound. After the reaction is complete, unreacted components are removed (e.g., by washing) and any complexes formed will remain immobilized on the solid surface. Where the non-immobilized species is pre-labeled, the detection of label immobilized on the surface indicates that complexes were formed. Where the non-immobilized species is not pre-labeled, an indirect label can be used to detect complexes anchored on the surface; e.g., using a labeled antibody specific for the initially non-immobilized species (the antibody, in turn, can be directly labeled or indirectly labeled with, e.g., a labeled anti-Ig antibody). Depending upon the order of addition of reaction components, test compounds that inhibit complex formation or that disrupt preformed complexes can be detected.

[0182] Alternatively, the reaction can be conducted in a liquid phase in the presence or absence of the test compound, the reaction products separated from unreacted components, and complexes detected; e.g., using an immobilized antibody specific for one of the binding components to anchor any complexes formed in solution, and a labeled antibody specific for the other partner to detect anchored complexes. Again, depending upon the order of addition of reactants to the liquid phase, test compounds that inhibit complex or that disrupt preformed complexes can be identified.

[0183] In an alternate embodiment, a homogeneous assay can be used. For example, a preformed complex of the target gene product and the interactive cellular or extracellular binding partner product is prepared in that either the target gene products or their binding partners are labeled, but the signal generated by the label is quenched due to complex formation (see, e.g., U.S. Patent No. 4,109,496 that utilizes this approach for immunoassays). The addition of a test substance that competes with and displaces one of the species from the preformed complex will result in the generation of a signal above background. In this way, test substances that disrupt target gene product-binding partner interaction can be identified.

[0184] Also, binding partners of *DLG1, KIAA0783, DPF3* or *CENPC1* molecules can be identified in a two-hybrid assay or three-hybrid assay (see, e.g., U.S. Patent No. 5,283,317; Zervos et al., Cell 72:223-232 (1993); Madura et al., J. Biol. Chem. 268: 12046-12054 (1993); Bartel et al., Biotechniques 14: 920-924 (1993); Iwabuchi et al., Oncogene 8: 1693-1696 (1993); and Brent WO94/10300), to identify other polypeptides, which bind to or interact with *DLG1, KIAA0783, DPF3* or *CENPC1* ("*DLG1, KIAA0783, DPF3* or *CENPC1*-binding polypeptides" or "*DLG1, KIAA0783, DPF3* or *CENPC1*-bp") and are involved in *DLG1, KIAA0783, DPF3* or *CENPC1* activity. Such *DLG1, KIAA0783, DPF3* or *CENPC1*-bps can be activators or inhibitors of signals by the *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptides or *DLG1, KIAA0783, DPF3* or *CENPC1* targets as, for example, downstream elements of a *DLG1, KIAA0783, DPF3* or *CENPC1*-mediated signaling pathway.

[0185] A two-hybrid system is based on the modular nature of most transcription factors, which consist of separable DNA-binding and activation domains. Briefly, the assay utilizes two different DNA constructs. In one construct, the gene that codes for a *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide is fused to a gene encoding the DNA binding domain of a known transcription factor (e.g., GAL-4). In the other construct, a DNA sequence, from a library of DNA sequences, that encodes an unidentified polypeptide ("prey" or "sample") is fused to a gene that codes for the activation domain of the known transcription factor. (Alternatively the: *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide can be the fused to the activator domain.) If the "bait" and the "prey" polypeptides are able to interact, in vivo, forming a *DLG1, KIAA0783, DPF3* or *CENPC1*-dependent complex, the DNA-binding and activation domains of the transcription factor are brought into close proximity. This proximity allows transcription

of a reporter gene (e.g., LacZ) which is operably linked to a transcriptional regulatory site responsive to the transcription factor. Expression of the reporter gene can be detected and cell colonies containing the functional transcription factor can be isolated and used to obtain the cloned gene which encodes the polypeptide which interacts with the *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide.

[0186] Candidate therapeutics for treating breast cancer are identified from a group of test molecules that interact with a *DLG1, KIAA0783, DPF3* or *CENPC1* nucleic acid or polypeptide. Test molecules are normally ranked according to the degree with which they interact or modulate (e.g., agonize or antagonize) DNA replication and/or processing, RNA transcription and/or processing, polypeptide production and/or processing, and/or function of *DLG1, KIAA0783, DPF3* or *CENPC1* molecules, for example, and then top ranking modulators are selected. In a preferred embodiment, the candidate therapeutic (i.e., test molecule) acts as a *DLG1, KIAA0783, DPF3* or *CENPC1* antagonist. Also, pharmacogenomic information described herein can determine the rank of a modulator. Candidate therapeutics typically are formulated for administration to a subject.

Therapeutic Treatments

[0187] Formulations or pharmaceutical compositions typically include in combination with a pharmaceutically acceptable carrier, a compound, an antisense nucleic acid, a ribozyme, an antibody, a binding partner that interacts with a *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide, a *DLG1, KIAA0783, DPF3* or *CENPC1* nucleic acid, or a fragment thereof. The formulated molecule may be one that is identified by a screening method described above. Also, formulations may comprise a *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide or fragment thereof. As used herein, the term "pharmaceutically acceptable carrier" includes solvents, dispersion media, coatings, antibacterial and antifungal agents, isotonic and absorption delaying agents, and the like, compatible with pharmaceutical administration. Supplementary active compounds can also be incorporated into the compositions.

[0188] A pharmaceutical composition is formulated to be compatible with its intended route of administration. Examples of routes of administration include parenteral, e.g., intravenous, intradermal, subcutaneous, oral (e.g., inhalation), transdermal (topical), transmucosal, and rectal administration. Solutions or suspensions used for parenteral, intradermal, or subcutaneous application can include the following components: a sterile diluent such as water for injection, saline solution, fixed oils, polyethylene glycols, glycerin, propylene glycol or other synthetic solvents; antibacterial agents such as benzyl alcohol or methyl parabens; antioxidants such as ascorbic acid or sodium bisulfite; chelating agents such as ethylenediaminetetraacetic acid; buffers such as acetates, citrates or phosphates and agents for the adjustment of tonicity such as sodium chloride or dextrose. pH can be adjusted with acids or

bases, such as hydrochloric acid or sodium hydroxide. The parenteral preparation can be enclosed in ampoules, disposable syringes or multiple dose vials made of glass or plastic.

[0189] Oral compositions generally include an inert diluent or an edible carrier. For the purpose of oral therapeutic administration, the active compound can be incorporated with excipients and used in the form of tablets, troches, or capsules, e.g., gelatin capsules. Oral compositions can also be prepared using a fluid carrier for use as a mouthwash. Pharmaceutically compatible binding agents, and/or adjuvant materials can be included as part of the composition. The tablets, pills, capsules, troches and the like can contain any of the following ingredients, or compounds of a similar nature: a binder such as microcrystalline cellulose, gum tragacanth or gelatin; an excipient such as starch or lactose, a disintegrating agent such as alginic acid, Primogel, or corn starch; a lubricant such as magnesium stearate or Sterotes; a glidant such as colloidal silicon dioxide; a sweetening agent such as sucrose or saccharin; or a flavoring agent such as peppermint, methyl salicylate, or orange flavoring.

[0190] Pharmaceutical compositions suitable for injectable use include sterile aqueous solutions (where water soluble) or dispersions and sterile powders for the extemporaneous preparation of sterile injectable solutions or dispersion. For intravenous administration, suitable carriers include physiological saline, bacteriostatic water, Cremophor EL™ (BASF, Parsippany, NJ) or phosphate buffered saline (PBS). In all cases, the composition must be sterile and should be fluid to the extent that easy syringability exists. It should be stable under the conditions of manufacture and storage and must be preserved against the contaminating action of microorganisms such as bacteria and fungi. The carrier can be a solvent or dispersion medium containing, for example, water, ethanol, polyol (for example, glycerol, propylene glycol, and liquid polyethylene glycol, and the like), and suitable mixtures thereof. The proper fluidity can be maintained, for example, by the use of a coating such as lecithin, by the maintenance of the required particle size in the case of dispersion and by the use of surfactants. Prevention of the action of microorganisms can be achieved by various antibacterial and antifungal agents, for example, parabens, chlorobutanol, phenol, ascorbic acid, thimerosal, and the like. In many cases, isotonic agents, for example, sugars, polyalcohols such as mannitol, sorbitol, sodium chloride sometimes are included in the composition. Prolonged absorption of the injectable compositions can be brought about by including in the composition an agent which delays absorption, for example, aluminum monostearate and gelatin.

[0191] Sterile injectable solutions can be prepared by incorporating the active compound in the required amount in an appropriate solvent with one or a combination of ingredients enumerated above, as required, followed by filtered sterilization. Generally, dispersions are prepared by incorporating the active compound into a sterile vehicle which contains a basic dispersion medium and the required other ingredients from those enumerated above. In the case of sterile powders for the preparation of sterile injectable solutions, methods of preparation often utilized are vacuum drying and freeze-drying which

yields a powder of the active ingredient plus any additional desired ingredient from a previously sterile-filtered solution thereof.

[0192] For administration by inhalation, the compounds are delivered in the form of an aerosol spray from pressured container or dispenser which contains a suitable propellant, e.g., a gas such as carbon dioxide, or a nebulizer.

[0193] Systemic administration can also be by transmucosal or transdermal means. For transmucosal or transdermal administration, penetrants appropriate to the barrier to be permeated are used in the formulation. Such penetrants are generally known in the art, and include, for example, for transmucosal administration, detergents, bile salts, and fusidic acid derivatives. Transmucosal administration can be accomplished through the use of nasal sprays or suppositories. For transdermal administration, the active compounds are formulated into ointments, salves, gels, or creams as generally known in the art. Molecules can also be prepared in the form of suppositories (e.g., with conventional suppository bases such as cocoa butter and other glycerides) or retention enemas for rectal delivery.

[0194] In one embodiment, active molecules are prepared with carriers that will protect the compound against rapid elimination from the body, such as a controlled release formulation, including implants and microencapsulated delivery systems. Biodegradable, biocompatible polymers can be used, such as ethylene vinyl acetate, polyanhydrides, polyglycolic acid, collagen, polyorthoesters, and polylactic acid. Methods for preparation of such formulations will be apparent to those skilled in the art. Materials can also be obtained commercially from Alza Corporation and Nova Pharmaceuticals, Inc. Liposomal suspensions (including liposomes targeted to infected cells with monoclonal antibodies to viral antigens) can also be used as pharmaceutically acceptable carriers. These can be prepared according to methods known to those skilled in the art, for example, as described in U.S. Patent No. 4,522,811.

[0195] It is advantageous to formulate oral or parenteral compositions in dosage unit form for ease of administration and uniformity of dosage. Dosage unit form as used herein refers to physically discrete units suited as unitary dosages for the subject to be treated; each unit containing a predetermined quantity of active compound calculated to produce the desired therapeutic effect in association with the required pharmaceutical carrier.

[0196] Toxicity and therapeutic efficacy of such compounds can be determined by standard pharmaceutical procedures in cell cultures or experimental animals, e.g., for determining the LD_{50} (the dose lethal to 50% of the population) and the ED_{50} (the dose therapeutically effective in 50% of the population). The dose ratio between toxic and therapeutic effects is the therapeutic index and it can be expressed as the ratio LD_{50}/ED_{50}. Molecules which exhibit high therapeutic indices often are utilized. While molecules that exhibit toxic side effects may be used, care should be taken to design a delivery

system that targets such compounds to the site of affected tissue in order to minimize potential damage to uninfected cells and, thereby, reduce side effects.

[0197] The data obtained from the cell culture assays and animal studies can be used in formulating a range of dosage for use in humans. The dosage of such molecules often lies within a range of circulating concentrations that include the ED_{50} with little or no toxicity. The dosage may vary within this range depending upon the dosage form employed and the route of administration utilized. For any molecules used in the methods described herein, the therapeutically effective dose can be estimated initially from cell culture assays. A dose may be formulated in animal models to achieve a circulating plasma concentration range that includes the IC_{50} (i.e., the concentration of the test compound which achieves a half-maximal inhibition of symptoms) as determined in cell culture. Such information can be used to more accurately determine useful doses in humans. Levels in plasma may be measured, for example, by high performance liquid chromatography.

[0198] As defined herein, a therapeutically effective amount of protein or polypeptide (i.e., an effective dosage) ranges from about 0.001 to 30 mg/kg body weight, sometimes about 0.01 to 25 mg/kg body weight, often about 0.1 to 20 mg/kg body weight, and more often about 1 to 10 mg/kg, 2 to 9 mg/kg, 3 to 8 mg/kg, 4 to 7 mg/kg, or 5 to 6 mg/kg body weight. The protein or polypeptide can be administered one time per week for between about 1 to 10 weeks, sometimes between 2 to 8 weeks, often between about 3 to 7 weeks, and more often for about 4, 5, or 6 weeks. The skilled artisan will appreciate that certain factors may influence the dosage and timing required to effectively treat a subject, including but not limited to the severity of the disease or disorder, previous treatments, the general health and/or age of the subject, and other diseases present. Moreover, treatment of a subject with a therapeutically effective amount of a protein, polypeptide, or antibody can include a single treatment, or sometimes can include a series of treatments.

[0199] With regard to polypeptide formulations, featured herein is a method for treating breast cancer in a subject, which comprises contacting one or more cells in the subject with a first *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide, where the subject comprises a second *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide having one or more polymorphic variations associated with cancer, and where the first polypeptide comprises fewer polymorphic variations associated with cancer than the second polypeptide. The first and second polypeptides are encoded by a nucleic acid which comprises a nucleotide sequence selected from the group consisting of the nucleotide sequence of SEQ ID NO: 1-8; a nucleotide sequence which encodes a polypeptide consisting of an amino acid sequence encoded by a nucleotide sequence of SEQ ID NO: 1-8; a nucleotide sequence which encodes a polypeptide that is 90% or more identical to an amino acid sequence encoded by a nucleotide sequence of SEQ ID NO: 1-8 and a

nucleotide sequence 90% or more identical to a nucleotide sequence of SEQ ID NO: 1-8. The subject is often a human.

[0200] For antibodies, a dosage of 0.1 mg/kg of body weight (generally 10 mg/kg to 20 mg/kg) is often utilized. If the antibody is to act in the brain, a dosage of 50 mg/kg to 100 mg/kg is often appropriate. Generally, partially human antibodies and fully human antibodies have a longer half-life within the human body than other antibodies. Accordingly, lower dosages and less frequent administration is often possible. Modifications such as lipidation can be used to stabilize antibodies and to enhance uptake and tissue penetration (e.g., into the brain). A method for lipidation of antibodies is described by Cruikshank et al., J. Acquired Immune Deficiency Syndromes and Human Retrovirology 14:193 (1997).

[0201] Antibody conjugates can be used for modifying a given biological response, the drug moiety is not to be construed as limited to classical chemical therapeutic agents. For example, the drug moiety may be a protein or polypeptide possessing a desired biological activity. Such proteins may include, for example, a toxin such as abrin, ricin A, pseudomonas exotoxin, or diphtheria toxin; a polypeptide such as tumor necrosis factor, .alpha.-interferon, .beta.-interferon, nerve growth factor, platelet derived growth factor, tissue plasminogen activator; or, biological response modifiers such as, for example, lymphokines, interleukin-1 ("IL-1"), interleukin-2 ("IL-2"), interleukin-6 ("IL-6"), granulocyte macrophage colony stimulating factor ("GM-CSF"), granulocyte colony stimulating factor ("G-CSF"), or other growth factors. Alternatively, an antibody can be conjugated to a second antibody to form an antibody heteroconjugate as described by Segal in U.S. Patent No. 4,676,980.

[0202] For compounds, exemplary doses include milligram or microgram amounts of the compound per kilogram of subject or sample weight, for example, about 1 microgram per kilogram to about 500 milligrams per kilogram, about 100 micrograms per kilogram to about 5 milligrams per kilogram, or about 1 microgram per kilogram to about 50 micrograms per kilogram. It is understood that appropriate doses of a small molecule depend upon the potency of the small molecule with respect to the expression or activity to be modulated. When one or more of these small molecules is to be administered to an animal (e.g., a human) in order to modulate expression or activity of a polypeptide or nucleic acid described herein, a physician, veterinarian, or researcher may, for example, prescribe a relatively low dose at first, subsequently increasing the dose until an appropriate response is obtained. In addition, it is understood that the specific dose level for any particular animal subject will depend upon a variety of factors including the activity of the specific compound employed, the age, body weight, general health, gender, and diet of the subject, the time of administration, the route of administration, the rate of excretion, any drug combination, and the degree of expression or activity to be modulated.

[0203] *DLG1, KIAA0783, DPF3* or *CENPC1* nucleic acid molecules can be inserted into vectors and used in gene therapy methods for treating breast cancer. Featured herein is a method for treating breast cancer in a subject, which comprises contacting one or more cells in the subject with a first *DLG1, KIAA0783, DPF3* or *CENPC1* nucleic acid, where genomic DNA in the subject comprises a second *DLG1, KIAA0783, DPF3* or *CENPC1* nucleic acid comprising one or more polymorphic variations associated with breast cancer, and where the first *DLG1, KIAA0783, DPF3* or *CENPC1* nucleic acid comprises fewer polymorphic variations associated with breast cancer. The first and second nucleic acids typically comprise a nucleotide sequence selected from the group consisting of the nucleotide sequence of SEQ ID NO: 1-8; a nucleotide sequence which encodes a polypeptide consisting of an amino acid sequence encoded by a nucleotide sequence of SEQ ID NO: 1-8; a nucleotide sequence that is 90% or more identical to the nucleotide sequence of SEQ ID NO: 1-8, and a nucleotide sequence which encodes a polypeptide that is 90% or more identical to an amino acid sequence encoded by a nucleotide sequence of SEQ ID NO: 1-8. The subject often is a human.

[0204] Gene therapy vectors can be delivered to a subject by, for example, intravenous injection, local administration (see U.S. Patent 5,328,470) or by stereotactic injection (see e.g., Chen et al., (1994) Proc. Natl. Acad. Sci. USA 91:3054-3057). Pharmaceutical preparations of gene therapy vectors can include a gene therapy vector in an acceptable diluent, or can comprise a slow release matrix in which the gene delivery vehicle is imbedded. Alternatively, where the complete gene delivery vector can be produced intact from recombinant cells (e.g., retroviral vectors) the pharmaceutical preparation can include one or more cells which produce the gene delivery system. Examples of gene delivery vectors are described herein.

[0205] Pharmaceutical compositions can be included in a container, pack, or dispenser together with instructions for administration.

[0206] Pharmaceutical compositions of active ingredients can be administered by any of the paths described herein for therapeutic and prophylactic methods for treating breast cancer. With regard to both prophylactic and therapeutic methods of treatment, such treatments may be specifically tailored or modified, based on knowledge obtained from pharmacogenomic analyses described herein. As used herein, the term "treatment" is defined as the application or administration of a therapeutic agent to a patient, or application or administration of a therapeutic agent to an isolated tissue or cell line from a patient, who has a disease, a symptom of disease or a predisposition toward a disease, with the purpose to cure, heal, alleviate, relieve, alter, remedy, ameliorate, improve or affect the disease, the symptoms of disease or the predisposition toward disease. A therapeutic agent includes, but is not limited to, small molecules, peptides, antibodies, ribozymes and antisense oligonucleotides.

[0207] Administration of a prophylactic agent can occur prior to the manifestation of symptoms characteristic of the *DLG1, KIAA0783, DPF3* or *CENPC1* aberrance, such that a disease or disorder is prevented or, alternatively, delayed in its progression. Depending on the type of *DLG1, KIAA0783, DPF3* or *CENPC1* aberrance, for example, a *DLG1, KIAA0783, DPF3* or *CENPC1* molecule, *DLG1, KIAA0783, DPF3* or *CENPC1* agonist, or *DLG1, KIAA0783, DPF3* or *CENPC1* antagonist agent can be used for treating the subject. The appropriate agent can be determined based on screening assays described herein.

[0208] As discussed, successful treatment of *DLG1, KIAA0783, DPF3* or *CENPC1* disorders can be brought about by techniques that serve to inhibit the expression or activity of target gene products. For example, compounds (e.g., an agent identified using an assays described above) that exhibit negative modulatory activity can be used to prevent and/or treat breast cancer. Such molecules can include, but are not limited to peptides, phosphopeptides, small organic or inorganic molecules, or antibodies (including, for example, polyclonal, monoclonal, humanized, anti-idiotypic, chimeric or single chain antibodies, and FAb, F(ab')2 and FAb expression library fragments, scFV molecules, and epitope-binding fragments thereof).

[0209] Further, antisense and ribozyme molecules that inhibit expression of the target gene can also be used to reduce the level of target gene expression, thus effectively reducing the level of target gene activity. Still further, triple helix molecules can be utilized in reducing the level of target gene activity. Antisense, ribozyme and triple helix molecules are discussed above.

[0210] It is possible that the use of antisense, ribozyme, and/or triple helix molecules to reduce or inhibit mutant gene expression can also reduce or inhibit the transcription (triple helix) and/or translation (antisense, ribozyme) of mRNA produced by normal target gene alleles, such that the concentration of normal target gene product present can be lower than is necessary for a normal phenotype. In such cases, nucleic acid molecules that encode and express target gene polypeptides exhibiting normal target gene activity can be introduced into cells via gene therapy method. Alternatively, in instances where the target gene encodes an extracellular polypeptide, normal target gene polypeptide often is co-administered into the cell or tissue to maintain the requisite level of cellular or tissue target gene activity.

[0211] Another method by which nucleic acid molecules may be utilized in treating or preventing a disease characterized by *DLG1, KIAA0783, DPF3* or *CENPC1* expression is through the use of aptamer molecules specific for *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide. Aptamers are nucleic acid molecules having a tertiary structure which permits them to specifically bind to polypeptide ligands (see, e.g., Osborne, et al., Curr. Opin. Chem. Biol.1(1): 5-9 (1997); and Patel, D. J., Curr. Opin. Chem. Biol. Jun;1(1): 32-46 (1997)). Since nucleic acid molecules may in many cases be more conveniently introduced into target cells than therapeutic polypeptide molecules may be, aptamers offer a method by

which *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide activity may be specifically decreased without the introduction of drugs or other molecules which may have pluripotent effects.

[0212] Antibodies can be generated that are both specific for target gene product and that reduce target gene product activity. Such antibodies may, therefore, by administered in instances whereby negative modulatory techniques are appropriate for the treatment of *DLG1, KIAA0783, DPF3* or *CENPC1* disorders. For a description of antibodies, see the Antibody section above.

[0213] In circumstances where injection of an animal or a human subject with a *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide or epitope for stimulating antibody production is harmful to the subject, it is possible to generate an immune response against *DLG1, KIAA0783, DPF3* or *CENPC1* through the use of anti-idiotypic antibodies (see, for example, Herlyn, D., Ann. Med.;31(1): 66-78 (1999); and Bhattacharya-Chatterjee & Foon, Cancer Treat. Res.; 94: 51-68 (1998)). If an anti-idiotypic antibody is introduced into a mammal or human subject, it should stimulate the production of anti-anti-idiotypic antibodies, which should be specific to the *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide. Vaccines directed to a disease characterized by *DLG1, KIAA0783, DPF3* or *CENPC1* expression may also be generated in this fashion.

[0214] In instances where the target antigen is intracellular and whole antibodies are used, internalizing antibodies may be utilized. Lipofectin or liposomes can be used to deliver the antibody or a fragment of the Fab region that binds to the target antigen into cells. Where fragments of the antibody are used, the smallest inhibitory fragment that binds to the target antigen often is utilized. For example, peptides having an amino acid sequence corresponding to the Fv region of the antibody can be used. Alternatively, single chain neutralizing antibodies that bind to intracellular target antigens can also be administered. Such single chain antibodies can be administered, for example, by expressing nucleotide sequences encoding single-chain antibodies within the target cell population (see e.g., Marasco et al., Proc. Natl. Acad. Sci. USA 90: 7889-7893 (1993)).

[0215] *DLG1, KIAA0783, DPF3* or *CENPC1* molecules and compounds that inhibit target gene expression, synthesis and/or activity can be administered to a patient at therapeutically effective doses to prevent, treat or ameliorate *DLG1, KIAA0783, DPF3* or *CENPC1* disorders. A therapeutically effective dose refers to that amount of the compound sufficient to result in amelioration of symptoms of the disorders.

[0216] Toxicity and therapeutic efficacy of such compounds can be determined by standard pharmaceutical procedures in cell cultures or experimental animals, e.g., for determining the LD_{50} (the dose lethal to 50% of the population) and the ED_{50} (the dose therapeutically effective in 50% of the population). The dose ratio between toxic and therapeutic effects is the therapeutic index and it can be expressed as the ratio LD_{50}/ED_{50}. Compounds that exhibit large therapeutic indices often are utilized.

While compounds that exhibit toxic side effects can be used, care should be taken to design a delivery system that targets such compounds to the site of affected tissue in order to minimize potential damage to uninfected cells and, thereby, reduce side effects.

[0217] Data obtained from cell culture assays and animal studies can be used in formulating a range of dosage for use in humans. The dosage of such compounds often lies within a range of circulating concentrations that include the ED_{50} with little or no toxicity. The dosage can vary within this range depending upon the dosage form employed and the route of administration utilized. For any compound used in a method described herein, the therapeutically effective dose can be estimated initially from cell culture assays. A dose can be formulated in animal models to achieve a circulating plasma concentration range that includes the IC_{50} (i.e., the concentration of the test compound that achieves a half-maximal inhibition of symptoms) as determined in cell culture. Such information can be used to more accurately determine useful doses in humans. Levels in plasma can be measured, for example, by high performance liquid chromatography.

[0218] Another example of effective dose determination for an individual is the ability to directly assay levels of "free" and "bound" compound in the serum of the test subject. Such assays may utilize antibody mimics and/or "biosensors" that have been created through molecular imprinting techniques. The compound which is able to modulate *DLG1, KIAA0783, DPF3* or *CENPC1* activity is used as a template, or "imprinting molecule", to spatially organize polymerizable monomers prior to their polymerization with catalytic reagents. The subsequent removal of the imprinted molecule leaves a polymer matrix which contains a repeated "negative image" of the compound and is able to selectively rebind the molecule under biological assay conditions. A detailed review of this technique can be seen in Ansell et al., Current Opinion in Biotechnology 7: 89-94 (1996) and in Shea, Trends in Polymer Science 2: 166-173 (1994). Such "imprinted" affinity matrixes are amenable to ligand-binding assays, whereby the immobilized monoclonal antibody component is replaced by an appropriately imprinted matrix. An example of the use of such matrixes in this way can be seen in Vlatakis, et al., Nature 361: 645-647 (1993). Through the use of isotope-labeling, the "free" concentration of compound which modulates the expression or activity of *DLG1, KIAA0783, DPF3* or *CENPC1* can be readily monitored and used in calculations of IC_{50}. Such "imprinted" affinity matrixes can also be designed to include fluorescent groups whose photon-emitting properties measurably change upon local and selective binding of target compound. These changes can be readily assayed in real time using appropriate fiberoptic devices, in turn allowing the dose in a test subject to be quickly optimized based on its individual IC_{50}. A rudimentary example of such a "biosensor" is discussed in Kriz et al., Analytical Chemistry 67: 2142-2144 (1995).

[0219] Provided herein are methods of modulating *DLG1, KIAA0783, DPF3* or *CENPC1* expression or activity for therapeutic purposes. Accordingly, in an exemplary embodiment, the modulatory method involves contacting a cell with a *DLG1, KIAA0783, DPF3* or *CENPC1* or agent that modulates one or more of the activities of *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide activity associated with the cell. An agent that modulates *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide activity can be an agent as described herein, such as a nucleic acid or a polypeptide, a naturally-occurring target molecule of a *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide (e.g., a *DLG1, KIAA0783, DPF3* or *CENPC1* substrate or receptor), a *DLG1, KIAA0783, DPF3* or *CENPC1* antibody, a *DLG1, KIAA0783, DPF3* or *CENPC1* agonist or antagonist, a peptidomimetic of a *DLG1, KIAA0783, DPF3* or *CENPC1* agonist or antagonist, or other small molecule.

[0220] In one embodiment, the agent stimulates one or more *DLG1, KIAA0783, DPF3* or *CENPC1* activities. Examples of such stimulatory agents include active *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide and a nucleic acid molecule encoding *DLG1, KIAA0783, DPF3* or *CENPC1*. In another embodiment, the agent inhibits one or more *DLG1, KIAA0783, DPF3* or *CENPC1* activities. Examples of such inhibitory agents include antisense *DLG1, KIAA0783, DPF3* or *CENPC1* nucleic acid molecules, anti-*DLG1, KIAA0783, DPF3* or *CENPC1* antibodies, and *DLG1, KIAA0783, DPF3* or *CENPC1* inhibitors. These modulatory methods can be performed in vitro (e.g., by culturing the cell with the agent) or, alternatively, in vivo (e.g., by administering the agent to a subject). As such, provided are methods of treating an individual afflicted with a disease or disorder characterized by aberrant or unwanted expression or activity of a *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide or nucleic acid molecule. In one embodiment, the method involves administering an agent (e.g., an agent identified by a screening assay described herein), or combination of agents that modulates (e.g., upregulates or downregulates) *DLG1, KIAA0783, DPF3* or *CENPC1* expression or activity. In a preferred embodiment, the method involves administering an agent (e.g., an agent identified by a screening assay described herein), or combination of agents that inhibits *DLG1, KIAA0783, DPF3* or *CENPC1* expression or activity. In another embodiment, the method involves administering a *DLG1, KIAA0783, DPF3* or *CENPC1* polypeptide or nucleic acid molecule as therapy to compensate for reduced, aberrant, or unwanted *DLG1, KIAA0783, DPF3* or *CENPC1* expression or activity.

[0221] Stimulation of *DLG1, KIAA0783, DPF3* or *CENPC1* activity is desirable in situations in which *DLG1, KIAA0783, DPF3* or *CENPC1* is abnormally downregulated and/or in which increased *DLG1, KIAA0783, DPF3* or *CENPC1* activity is likely to have a beneficial effect. For example, stimulation of *DLG1, KIAA0783, DPF3* or *CENPC1* activity is desirable in situations in which a *DLG1, KIAA0783, DPF3* or *CENPC1* is downregulated and/or in which increased *DLG1, KIAA0783, DPF3* or *CENPC1* activity is likely to have a beneficial effect. Likewise, inhibition of *DLG1, KIAA0783, DPF3*

or *CENPC1* activity is desirable in situations in which *DLG1, KIAA0783, DPF3* or *CENPC1* is abnormally upregulated and/or in which decreased *DLG1, KIAA0783, DPF3* or *CENPC1* activity is likely to have a beneficial effect.

Methods of Treatment

[0222] In another aspect, provided are methods for identifying a risk of cancer in an individual as described herein and, if a genetic predisposition is identified, treating that individual to delay or reduce or prevent the development of cancer. Such a procedure can be used to treat breast cancer. Optionally, treating an individual for cancer may include inhibiting cellular proliferation, inhibiting metastasis, inhibiting invasion, or preventing tumor formation or growth as defined herein. Suitable treatments to prevent or reduce or delay breast cancer focus on inhibiting additional cellular proliferation, inhibiting metastasis, inhibiting invasion, and preventing further tumor formation or growth. Treatment usually includes surgery followed by radiation therapy. Surgery may be a lumpectomy or a mastectomy (e.g., total, simple or radical). Even if the doctor removes all of the cancer that can be seen at the time of surgery, the patient may be given radiation therapy, chemotherapy, or hormone therapy after surgery to try to kill any cancer cells that may be left. Radiation therapy is the use of x-rays or other types of radiation to kill cancer cells and shrink tumors. Radiation therapy may use external radiation (using a machine outside the body) or internal radiation. Chemotherapy is the use of drugs to kill cancer cells. Chemotherapy may be taken by mouth, or it may be put into the body by inserting a needle into a vein or muscle. Hormone therapy often focuses on estrogen and progesterone, which are hormones that affect the way some cancers grow. If tests show that the cancer cells have estrogen and progesterone receptors (molecules found in some cancer cells to which estrogen and progesterone will attach), hormone therapy is used to block the way these hormones help the cancer grow. Hormone therapy with tamoxifen is often given to patients with early stages of breast cancer and those with metastatic breast cancer. Other types of treatment being tested in clinical trials include sentinel lymph node biopsy followed by surgery and high-dose chemotherapy with bone marrow transplantation and peripheral blood stem cell transplantation. Any preventative/therapeutic treatment known in the art may be prescribed and/or administered, including, for example, surgery, chemotherapy and/or radiation treatment, and any of the treatments may be used in combination with one another to treat or prevent breast cancer (e.g., surgery followed by radiation therapy).

[0223] Also provided are methods of preventing or treating cancer comprising providing an individual in need of such treatment with a *DLG1, KIAA0783, DPF3* or *CENPC1* inhibitor that reduces or inhibits the overexpression of mutant *DLG1, KIAA0783, DPF3* or *CENPC1* (e.g., a *DLG1, KIAA0783, DPF3* or *CENPC1* polynucleotide with an allele that is associated with cancer). Included herein are

methods of reducing or blocking the expression of *DLG1, KIAA0783, DPF3* or *CENPC1* comprising providing or administering to individuals in need of reducing or blocking the expression of *DLG1, KIAA0783, DPF3* or *CENPC1* a pharmaceutical or physiologically acceptable composition comprising a molecule capable of inhibiting expression of *DLG1, KIAA0783, DPF3* or *CENPC1*, e.g., a siRNA molecule. Also included herein are methods of reducing or blocking the expression of secondary regulatory genes regulated by *DLG1, KIAA0783, DPF3* or *CENPC1* that play a role in oncogenesis which comprises introducing competitive inhibitors that target *DLG1, KIAA0783, DPF3* or *CENPC1*'s effect on these regulatory genes or that block the binding of positive factors necessary for the expression of these regulatory genes.

[0224] The examples set forth below are intended to illustrate but not limit the invention.

Examples

[0225] In the following studies a group of subjects were selected according to specific parameters relating to breast cancer. Nucleic acid samples obtained from individuals in the study group were subjected to genetic analysis, which identified associations between breast cancer and certain polymorphic regions in the *DLG1, KIAA0783, DPF3* and *CENPC1* genes (herein referred to as "target genes", "target nucleotides", "target polypeptides" or simply "targets"). Methods are described for producing DLG1, KIAA0783, DPF3 and CENPC1 polypeptides and polypeptide variants *in vitro* or *in vivo*. DLG1, KIAA0783, DPF3 and CENPC1 nucleic acids or polypeptides and variants thereof are utilized for screening test molecules for those that interact with DLG1, KIAA0783, DPF3 and CENPC1 molecules. Test molecules identified as interactors with DLG1, KIAA0783, DPF3 and CENPC1 molecules and variants are further screened *in vivo* to determine whether they treat breast cancer.

Example 1
Samples and Pooling Strategies

Sample Selection

[0226] Blood samples were collected from individuals diagnosed with breast cancer, which were referred to as case samples. Also, blood samples were collected from individuals not diagnosed with breast cancer as gender and age-matched controls. All of the samples were of German/German descent. A database was created that listed all phenotypic trait information gathered from individuals for each case and control sample. Genomic DNA was extracted from each of the blood samples for genetic analyses.

DNA Extraction from Blood Samples

[0227] Six to ten milliliters of whole blood was transferred to a 50 ml tube containing 27 ml of red cell lysis solution (RCL). The tube was inverted until the contents were mixed. Each tube was incubated for 10 minutes at room temperature and inverted once during the incubation. The tubes were then centrifuged for 20 minutes at 3000 x g and the supernatant was carefully poured off. 100-200 µl of residual liquid was left in the tube and was pipetted repeatedly to resuspend the pellet in the residual supernatant. White cell lysis solution (WCL) was added to the tube and pipetted repeatedly until completely mixed. While no incubation was normally required, the solution was incubated at 37°C or room temperature if cell clumps were visible after mixing until the solution was homogeneous. 2 ml of protein precipitation was added to the cell lysate. The mixtures were vortexed vigorously at high speed for 20 sec to mix the protein precipitation solution uniformly with the cell lysate, and then centrifuged for 10 minutes at 3000 x g. The supernatant containing the DNA was then poured into a clean 15 ml tube, which contained 7 ml of 100% isopropanol. The samples were mixed by inverting the tubes gently until white threads of DNA were visible. Samples were centrifuged for 3 minutes at 2000 x g and the DNA was visible as a small white pellet. The supernatant was decanted and 5 ml of 70% ethanol was added to each tube. Each tube was inverted several times to wash the DNA pellet, and then centrifuged for 1 minute at 2000 x g. The ethanol was decanted and each tube was drained on clean absorbent paper. The DNA was dried in the tube by inversion for 10 minutes, and then 1000 µl of 1X TE was added. The size of each sample was estimated, and less TE buffer was added during the following DNA hydration step if the sample was smaller. The DNA was allowed to rehydrate overnight at room temperature, and DNA samples were stored at 2-8°C.

[0228] DNA was quantified by placing samples on a hematology mixer for at least 1 hour. DNA was serially diluted (typically 1:80, 1:160, 1:320, and 1:640 dilutions) so that it would be within the measurable range of standards. 125 µl of diluted DNA was transferred to a clear U-bottom microtitre plate, and 125 µl of 1X TE buffer was transferred into each well using a multichannel pipette. The DNA and 1X TE were mixed by repeated pipetting at least 15 times, and then the plates were sealed. 50 µl of diluted DNA was added to wells A5-H12 of a black flat bottom microtitre plate. Standards were inverted six times to mix them, and then 50 µl of 1X TE buffer was pipetted into well A1, 1000 ng/ml of standard was pipetted into well A2, 500 ng/ml of standard was pipetted into well A3, and 250 ng/ml of standard was pipetted into well A4. PicoGreen (Molecular Probes, Eugene, Oregon) was thawed and freshly diluted 1:200 according to the number of plates that were being measured. PicoGreen was vortexed and then 50µl was pipetted into all wells of the black plate with the diluted DNA. DNA and PicoGreen were mixed by pipetting repeatedly at least 10 times with the multichannel pipette. The plate was placed into a Fluoroskan Ascent Machine (microplate fluorometer produced by Labsystems) and the samples were

allowed to incubate for 3 minutes before the machine was run using filter pairs 485 nm excitation and 538 nm emission wavelengths. Samples having measured DNA concentrations of greater than 450 ng/μl were re-measured for conformation. Samples having measured DNA concentrations of 20 ng/μl or less were re-measured for confirmation.

Pooling Strategies

[0229] Samples were placed into one of two groups based on disease status. The two groups were female case groups and female control groups. A select set of samples from each group were utilized to generate pools, and one pool was created for each group. Each individual sample in a pool was represented by an equal amount of genomic DNA. For example, where 25 ng of genomic DNA was utilized in each PCR reaction and there were 200 individuals in each pool, each individual would provide 125 pg of genomic DNA. Inclusion or exclusion of samples for a pool was based upon the following criteria: the sample was derived from an individual characterized as Caucasian; the sample was derived from an individual of German paternal and maternal descent; the database included relevant phenotype information for the individual; case samples were derived from individuals diagnosed with breast cancer; control samples were derived from individuals free of cancer and no family history of breast cancer; and sufficient genomic DNA was extracted from each blood sample for all allelotyping and genotyping reactions performed during the study. Phenotype information included pre- or post-menopausal, familial predisposition, country or origin of mother and father, diagnosis with breast cancer (date of primary diagnosis, age of individual as of primary diagnosis, grade or stage of development, occurrence of metastases, e.g., lymph node metastases, organ metastases), condition of body tissue (skin tissue, breast tissue, ovary tissue, peritoneum tissue and myometrium), method of treatment (surgery, chemotherapy, hormone therapy, radiation therapy). Samples that met these criteria were added to appropriate pools based on gender and disease status.

[0230] The selection process yielded the pools set forth in Table 1, which were used in the studies that follow:

Table 1

	Female CASE	Female CONTROL
Pool size (Number)	272	276
Pool Criteria (ex: case/control)	case	control
Mean Age (ex: years)	59.6	55.4

Example 2

Association of Polymorphic Variants with Breast cancer

[0231] A whole-genome screen was performed to identify particular SNPs associated with occurrence of breast cancer. As described in Example 1, two sets of samples were utilized, which included samples from female individuals having breast cancer (breast cancer cases) and samples from female individuals not having cancer (female controls). The initial screen of each pool was performed in an allelotyping study, in which certain samples in each group were pooled. By pooling DNA from each group, an allele frequency for each SNP in each group was calculated. These allele frequencies were then compared to one another. Particular SNPs were considered as being associated with breast cancer when allele frequency differences calculated between case and control pools were statistically significant. SNP disease association results obtained from the allelotyping study were then validated by genotyping each associated SNP across all samples from each pool. The results of the genotyping were then analyzed, allele frequencies for each group were calculated from the individual genotyping results, and a p-value was calculated to determine whether the case and control groups had statistically significantly differences in allele frequencies for a particular SNP. When the genotyping results agreed with the original allelotyping results, the SNP disease association was considered validated at the genetic level.

SNP Panel Used for Genetic Analyses

[0232] A whole-genome SNP screen began with an initial screen of approximately 25,000 SNPs over each set of disease and control samples using a pooling approach. The pools studied in the screen are described in Example 1. The SNPs analyzed in this study were part of a set of 25,488 SNPs confirmed as being statistically polymorphic as each is characterized as having a minor allele frequency of greater than 10%. The SNPs in the set reside in genes or in close proximity to genes, and many reside in gene exons. Specifically, SNPs in the set are located in exons, introns, and within 5,000 base-pairs upstream of a transcription start site of a gene. In addition, SNPs were selected according to the following criteria: they are located in ESTs; they are located in Locuslink or Ensemble genes; and they are located in Genomatix promoter predictions. SNPs in the set also were selected on the basis of even spacing across the genome, as depicted in Table 2.

[0233] A case-control study design using a whole genome association strategy involving approximately 28,000 single nucleotide polymorphisms (SNPs) was employed. Approximately 25,000 SNPs were evenly spaced in gene-based regions of the human genome with a median inter-marker distance of about 40,000 base pairs. Additionally, approximately 3,000 SNPs causing amino acid substitutions in genes described in the literature as candidates for various diseases were used. The case-control study samples were of female German origin (German paternal and maternal descent) 548

individuals were equally distributed in two groups (female controls and female cases). The whole genome association approach was first conducted on 2 DNA pools representing the 2 groups. Significant markers were confirmed by individual genotyping.

Table 2

General Statistics		Spacing Statistics	
Total # of SNPs	25,488	Median	37,058 bp
# of Exonic SNPs	>4,335 (17%)	Minimum*	1,000 bp
# SNPs with refSNP ID	20,776 (81%)	Maximum*	3,000,000 bp
Gene Coverage	>10,000	Mean	122,412 bp
Chromosome Coverage	All	Std Deviation	373,325 bp
		**Excludes outliers*	

Allelotyping and Genotyping Results

[0234] The genetic studies summarized above and described in more detail below identified allelic variants associated with breast cancer. The allelic variants identified from the SNP panel described in Table 2 are summarized below in Table 3.

Table 3

SNP Reference	Chromo-some Position	Position in Figs 1-4	Contig Identification	Contig Position	Sequence Identification	Sequence Position	Allelic Variability
rs1949471	198272877	39977	NT_029928	1484976	NM_004087	Exonic (R278Q)	T/C
rs220097	10793860	49860	NT_007819	10345196	NM_014660	intragenic	T/C
rs1990440	71267195	40095	NT_026437	53197195	NM_012074	intragenic	G/C
rs355510	68321769	46769	NT_022778	8587277	NM_001812	intragenic	G/A

[0235] Table 3 includes information pertaining to the incident polymorphic variant associated with breast cancer identified herein. Public information pertaining to the polymorphism and the genomic sequence that includes the polymorphism are indicated. The genomic sequences identified in Table 3 may be accessed at the http address www.ncbi.nih.gov/entrez/query.fcgi, for example, by using the publicly available SNP reference number (*e.g.*, rs1949471). The chromosome position refers to the position of the SNP within NCBI's Genome Build 33, which may be accessed at the following http address: www.ncbi.nlm.nih.gov/mapview/map_search.cgi?chr=hum_chr.inf&query=. The "Contig Position" provided in Table 3 corresponds to a nucleotide position set forth in the contig sequence, and designates the polymorphic site corresponding to the SNP reference number. The sequence containing the polymorphisms also may be referenced by the "Sequence Identification" set forth in Table 3. The "Sequence Identification" corresponds to cDNA sequence that encodes associated target polypeptides

(*e.g.*, *DLG1*) of the invention. The position of the SNP within the cDNA sequence is provided in the "Sequence Position" column of Table 3. Also, the allelic variation at the polymorphic site and the allelic variant identified as associated with breast cancer is specified in Table 3. All nucleotide sequences referenced and accessed by the parameters set forth in Table 3 are incorporated herein by reference. rs220097 also is known rs286246.

Assay for Verifying, Allelotyping, and Genotyping SNPs

[0236] A MassARRAY™ system (Sequenom, Inc.) was utilized to perform SNP genotyping in a high-throughput fashion. This genotyping platform was complemented by a homogeneous, single-tube assay method (hME™ or homogeneous MassEXTEND™ (Sequenom, Inc.)) in which two genotyping primers anneal to and amplify a genomic target surrounding a polymorphic site of interest. A third primer (the MassEXTEND™ primer), which is complementary to the amplified target up to but not including the polymorphism, was then enzymatically extended one or a few bases through the polymorphic site and then terminated.

[0237] For each polymorphism, SpectroDESIGNER™ software (Sequenom, Inc.) was used to generate a set of PCR primers and a MassEXTEND™ primer was used to genotype the polymorphism. Table 4 shows PCR primers and Table 5 shows extension primers used for analyzing polymorphisms. The initial PCR amplification reaction was performed in a 5 μl total volume containing 1X PCR buffer with 1.5 mM $MgCl_2$ (Qiagen), 200 μM each of dATP, dGTP, dCTP, dTTP (Gibco-BRL), 2.5 ng of genomic DNA, 0.1 units of HotStar DNA polymerase (Qiagen), and 200 nM each of forward and reverse PCR primers specific for the polymorphic region of interest.

Table 4: PCR Primers

Reference SNP ID	Forward PCR primer	Reverse PCR primer
rs1949471	ACGTTGGATGGCTTCAACTGCTTTGCTATG	ACGTTGGATGTTTCTCAGGGTCAATGACTG
rs220097	GCAAACGTGCACATTTGCAC	TTCCTGGGAATGGATTTCAG
rs1990440	CCAGGGTGTGTTCTAATACG	AAGTCACTAACCCCACACAC
rs355510	TTCTGAGATGATCCTGATGG	CCCTCCTTTTAACCTTTTAGG

[0238] Samples were incubated at 95°C for 15 minutes, followed by 45 cycles of 95°C for 20 seconds, 56°C for 30 seconds, and 72°C for 1 minute, finishing with a 3 minute final extension at 72°C. Following amplification, shrimp alkaline phosphatase (SAP) (0.3 units in a 2 μl volume) (Amersham Pharmacia) was added to each reaction (total reaction volume was 7 μl) to remove any residual dNTPs

that were not consumed in the PCR step. Samples were incubated for 20 minutes at 37°C, followed by 5 minutes at 85°C to denature the SAP.

[0239] Once the SAP reaction was complete, a primer extension reaction was initiated by adding a polymorphism-specific MassEXTEND™ primer cocktail to each sample. Each MassEXTEND™ cocktail included a specific combination of dideoxynucleotides (ddNTPs) and deoxynucleotides (dNTPs) used to distinguish polymorphic alleles from one another. In Table 5, ddNTPs are shown and the fourth nucleotide not shown is the dNTP.

Table 5: Extend Primers

Reference SNP ID	Extend Probe	Term Mix
rs1949471	CAGGGTCAATGACTGTATATTAC	ACT
rs220097	ACAGAGTTTTAAACCTCCTACA	ACT
rs1990440	CGTCAGCAAATGTGTACCGA	ACT
rs355510	ATGGTTTTCTTTCTTGTCCTTC	ACG

[0240] The MassEXTEND™ reaction was performed in a total volume of 9 µl, with the addition of 1X ThermoSequenase buffer, 0.576 units of ThermoSequenase (Amersham Pharmacia), 600 nM MassEXTEND™ primer, 2 mM of ddATP and/or ddCTP and/or ddGTP and/or ddTTP, and 2 mM of dATP or dCTP or dGTP or dTTP. The deoxy nucleotide (dNTP) used in the assay normally was complementary to the nucleotide at the polymorphic site in the amplicon. Samples were incubated at 94°C for 2 minutes, followed by 55 cycles of 5 seconds at 94°C, 5 seconds at 52°C, and 5 seconds at 72°C.

[0241] Following incubation, samples were desalted by adding 16 µl of water (total reaction volume was 25 µl), 3 mg of SpectroCLEAN™ sample cleaning beads (Sequenom, Inc.) and allowed to incubate for 3 minutes with rotation. Samples were then robotically dispensed using a piezoelectric dispensing device (SpectroJET™ (Sequenom, Inc.)) onto either 96-spot or 384-spot silicon chips containing a matrix that crystallized each sample (SpectroCHIP® (Sequenom, Inc.)). Subsequently, MALDI-TOF mass spectrometry (Biflex and Autoflex MALDI-TOF mass spectrometers (Bruker Daltonics) can be used) and SpectroTYPER RT™ software (Sequenom, Inc.) were used to analyze and interpret the SNP genotype for each sample.

Genetic Analysis

[0242] Variations identified in the target genes are provided in their respective genomic sequences (see Figures 1-5) Minor allelic frequencies for these polymorphisms was verified as being 10% or

greater by determining the allelic frequencies using the extension assay described above in a group of samples isolated from 92 individuals originating from the state of Utah in the United States, Venezuela and France (Coriell cell repositories).

[0243] Genotyping results are shown for female pools in Table 6A and 6B. Table 6A shows the orginal genotyping results and Table 6B shows the genotyped results re-analyzed to remove duplicate individuals from the cases and controls (*i.e.*, individuals who were erroneously included more than once as either cases or controls). Therefore, Table 6B represents a more accurate measure of the allele frequencies for this particular SNP. In the subsequent tables, "AF" refers to allelic frequency; and "F case" and "F control" refer to female case and female control groups, respectively.

Table 6A

Reference SNP ID	AF F case	AF F control	p-value	Odds Ratio	Breast Cancer Assoc. Allele
rs1949471	T = 0.186 C = 0.814	T = 0.112 C = 0.890	**0.0005**	0.54	T
rs220097	T = 0.721 C = 0.279	T = 0.626 C = 0.374	**0.0014**	0.66	T
rs1990440	C = 0.876 G = 0.124	C = 0.926 G = 0.074	**0.0027**	0.65	G
rs355510	A = 0.545 G = 0.455	A = 0.616 G = 0.384	**0.0173**	0.75	G

Table 6B

Reference SNP ID	AF F case	AF F control	p-value	Odds Ratio	Breast Cancer Assoc. Allele
rs1949471	T = 0.182 C = 0.818	T = 0.108 C = 0.892	**0.0009**	0.54	T
rs220097	T = 0.709 C = 0.291	T = 0.624 C = 0.376	**0.0045**	0.68	T
rs1990440	C = 0.879 G = 0.121	C = 0.915 G = 0.085	**0.0692**	0.67	G
rs355510	A = 0.539 G = 0.461	A = 0.617 G = 0.383	**0.0123**	0.73	G

[0244] The single marker alleles set forth in Table 3 were considered validated, since the genotyping data for the females, males or both pools were significantly associated with breast cancer, and because the genotyping results agreed with the original allelotyping results. Particularly significant associations with breast cancer are indicated by a calculated p-value of less than 0.05 for genotype results, which are set forth in bold text. Tables 6A and 6B show the disease associated allele in column 6. In the case of

rs1949471, this SNP is an exonic SNP that codes for a R278Q amino acid change in the DLG1 gene. The thymine allele codes for glutamine (Q); therefore, a glutamine is associated with an increased risk of breast cancer.

[0245] Odds ratio results are shown in Tables 6A and 6B. An odds ratio is an unbiased estimate of relative risk which can be obtained from most case-control studies. Relative risk (RR) is an estimate of the likelihood of disease in the exposed group (susceptibility allele or genotype carriers) compared to the unexposed group (not carriers). It can be calculated by the following equation:

RR = *I*A/*I*a

*I*A is the incidence of disease in the A carriers and *I*a is the incidence of disease in the non-carriers.

RR > 1 indicates the A allele increases disease susceptibility.

RR < 1 indicates the a allele increases disease susceptibility.

[0246] For example, RR = 1.5 indicates that carriers of the A allele have 1.5 times the risk of disease than non-carriers, *i.e.*, 50% more likely to get the disease.

[0247] Case-control studies do not allow the direct estimation of *I*A and *I*a, therefore relative risk cannot be directly estimated. However, the odds ratio (OR) can be calculated using the following equation:

OR = (nDAnda)/(ndAnDa) = *p*DA(1 – pdA)/*p*dA(1 – pDA), or

OR = ((case f) / (1- case f)) / ((control f) / (1-control f)), where f = susceptibility allele frequency.

[0248] An odds ratio can be interpreted in the same way a relative risk is interpreted and can be directly estimated using the data from case-control studies, *i.e.*, case and control allele frequencies. The higher the odds ratio value, the larger the effect that particular allele has on the development of breast cancer. Possessing an allele associated with a relatively high odds ratio translates to having a higher risk of developing or having breast cancer.

Example 3

DLG1 Region Proximal SNPs

[0249] It has been discovered that a polymorphic variation (rs1949471) in a region that encodes the discs, large homolog 1 (Drosophila) (DLG1) gene is associated with the occurrence of breast cancer (see Examples 1 and 2). Subsequently, SNPs proximal to the incident SNP (rs1949471) were identified and allelotyped in breast cancer sample sets and control sample sets as described in Examples 1 and 2. Approximately twenty-one allelic variants located within the DLG1 region were identified and allelotyped. The polymorphic variants are set forth in Table 7. The chromosome position provided in column four of Table 7 is based on Genome "Build 33" of NCBI's GenBank.

Table 7

dbSNP rs#	Chromosome	Chromosome Position	Position in Figure 1	Allele Variants
2341225	3	198233033	133	T/C
3856760	3	198240838	7938	T/C
2195027	3	198241773	8873	G/A
1356612	3	198246121	13221	C/T
3773845	3	198250188	17288	T/C
2098941	3	198258632	25732	G/A
890491	3	198259823	26923	C/G
1949471	3	198272877	39977	C/T
3773851	3	198274184	41284	T/A
3773852	3	198274310	41410	A/C
3773853	3	198274377	41477	C/T
1195059	3	198274414	41514	G/A
3773855	3	198275506	42606	G/A
3821713	3	198275642	42742	A/C
604005	3	198292415	59515	G/A
DLG1_SNP	3	198292708	59808	T/C
2879969	3	198293165	60265	C/G
958902	3	198300052	67152	T/C
1839742	3	198301232	68332	T/C
1868890	3	198304028	71128	T/C
1868891	3	198309327	76427	G/A

Assay for Verifying and Allelotyping SNPs

[0250] The methods used to verify and allelotype the proximal SNPs of Table 7 are the same methods described in Examples 1 and 2 herein. The PCR primers and extend primers used in these assays are provided in Table 8 and Table 9, respectively.

Table 8

dbSNP rs#	Forward PCR primer	Reverse PCR primer
604005	ACGTTGGATGTGTCTCGCTTTTAGCCTGTG	ACGTTGGATGCAGACAGACATACAGAAGGG
890491	ACGTTGGATGGCAGAACCATGGAGAAAAGC	ACGTTGGATGGGCAAGAGTAAGGCACTATC
958902	ACGTTGGATGGCCACTGAATTGTACATTAAC	ACGTTGGATGATTGGAGTCCCGAGCTAAAC
1195059	ACGTTGGATGCCTGTTTTCATTTAGACTCC	ACGTTGGATGTGCTCACAAAGATTTAAACC
1356612	ACGTTGGATGTTGAACAGCTCAGCTGAAAG	ACGTTGGATGAGATACATGTCTTGTCTGGG
1839742	ACGTTGGATGTCTGAGGTCAGGAGTTTGAG	ACGTTGGATGGCCACCATGTCCAGCTAATT
1868890	ACGTTGGATGAGTGAGGAAGGCCTATTAAC	ACGTTGGATGATACCTGAGTCGAACTCTTG
1868891	ACGTTGGATGTTATTGCTCTTGAACGTGGC	ACGTTGGATGTCTGAGAAAAAGAATTGGGG
1949471	ACGTTGGATGTTTCTCAGGGTCAATGACTG	ACGTTGGATGAGACCCTGCTTCTTTCAACG
2098941	ACGTTGGATGATTAGCTGGGCATGCTATCC	ACGTTGGATGTGTAGCCTTGAATTCCTGGG

dbSNP rs#	Forward PCR primer	Reverse PCR primer
2195027	ACGTTGGATGGGCGCTAAATAATGCGCCAC	ACGTTGGATGCTGACCTCGTGATCTGCCTG
2341225	ACGTTGGATGGGCGGGTGGGAAGACTCTAA	ACGTTGGATGTCTTTCACTGTATTCAGATC
2879969	ACGTTGGATGCTCCATTTCAAAAAAAAAAA	ACGTTGGATGCCTTAGAGGTATGTCCAGAG
3773845	ACGTTGGATGACACAAGTAACAAACTTGAG	ACGTTGGATGGTGCTTGAAGAAATTATGTG
3773851	ACGTTGGATGTAAGATACGGAGGATAGAGG	ACGTTGGATGGCATATAGTCTTTGTGGTGTG
3773852	ACGTTGGATGGTGAGTGTACTTAAATAAGTT	ACGTTGGATGGTTTCCCTTTGTGTTTTCAG
3773853	ACGTTGGATGTGGTTTAAATCTTTGTGAGC	ACGTTGGATGCTGTGAGTGTATCTGAAAAC
3773855	ACGTTGGATGGCTTGTTTTATGAACTGGAG	ACGTTGGATGTTAATACCATTGGTTAAATC
3821713	ACGTTGGATGTTCAGGCAACTCAAGTAAGC	ACGTTGGATGTAGAGTGGGTGTTTACACTG
3856760	ACGTTGGATGTGATCTCAGCTCACTGTAAC	ACGTTGGATGTGTAGTCCCAGCTACTCAGG
FCH-1723	ACGTTGGATGGCTTCAACTGCTTTGCTATG	ACGTTGGATGTTTCTCAGGGTCAATGACTG
DLG1_SNP	ACGTTGGATGCTTCATAGTAGCCAGGCTAG	ACGTTGGATGAGCACATGAACAGATGTGTC

Table 9

dbSNP rs#	Extend Primer	Term Mix
604005	TTATCAACCTACAATGGA	ACG
890491	TTATGGCCATACGTAAAAAGCA	ACT
958902	CGGAGGCTTTATTCGTA	ACT
1195059	AAAGATTTAAACCATCAACCAAAT	ACG
1356612	GGGTAGTGGTTTCATGATTTTTA	ACG
1839742	TCCAGCTAATTTTTTGTATTTTTA	ACT
1868890	CTGAGTCGAACTCTTGTATAAA	ACT
1868891	GAAAAAGAATTGGGGATTATAAC	ACG
1949471	CGAACATCTACTTCATTTACT	ACG
2098941	TCCTCCCACATCAGCCT	ACG
2195027	GCGTGAGCCACCACACC	ACG
2341225	CACTGTATTCAGATCTTCATATTT	ACT
2879969	CATCATACTGCCTCTGG	ACT
3773845	TTATGTGTTCTCTATTTATTGACT	ACT
3773851	TTTGTGGTGTGGGATTC	CGT
3773852	TATTTTCCATTTCCTCTCTG	ACT
3773853	AAGGGAAACTCATGATTTCTA	ACG
3773855	AGGCTTTTTGTAGCAGT	ACG
3821713	GTGGGTGTTTACACTGTTTAATAC	ACT
3856760	ATGAGAATCACTTGAACCTG	ACT
FCH-1723	CAGGGTCAATGACTGTATATTAC	ACT
DLG1_SNP	AGATGTGTCACAAATGCAA	ACT

Genetic Analysis of Allelotyping Results

[0251] Allelotyping results are shown for cases and controls in Table 10. The allele frequency for the A2 allele is noted in the fifth and sixth columns for breast cancer pools and control pools, respectively, where "AF" is allele frequency. The allele frequency for the A1 allele can be easily calculated by subtracting the A2 allele frequency from 1 (A1 AF = 1-A2 AF). For example, the SNP rs2341225 has the following case and control allele frequencies: case A1 (T) = 0.747; case A2 (C) = 0.253; control A1 (T) = 0.743; and control A2 (C) = 0.257, where the nucleotide is provided in paranthesis. SNPs with blank allele frequencies were untyped.

Table 10

dbSNP rs#	Chromosome	Position in Figure 1	Allele Variants	A2 Case AF	A2 Control AF	p-Value
2341225	198233033	133	T/C	0.253	0.257	0.8897
3856760	198240838	7938	T/C	0.959	0.985	0.0095
2195027	198241773	8873	G/A	0.651	0.691	0.1538
1356612	198246121	13221	C/T	0.197	0.243	0.0653
3773845	198250188	17288	T/C	0.415	0.414	0.9646
2098941	198258632	25732	G/A	0.281	0.335	0.0515
890491	198259823	26923	C/G	0.440	0.525	0.0051
1949471	198272877	39977	C/T	0.181	0.092	0.0001
3773851	198274184	41284	T/A	0.351	0.371	0.4824
3773852	198274310	41410	A/C	0.206	0.233	0.2786
3773853	198274377	41477	C/T	0.485	0.480	0.8660
1195059	198274414	41514	G/A	0.936	0.931	0.7361
3773855	198275506	42606	G/A	0.275	0.260	0.5723
3821713	198275642	42742	A/C	0.728	0.677	0.0666
604005	198292415	59515	G/A	0.985	0.986	0.8647
DLG1_SNP	198292708	59808	T/C	0.723	0.825	0.0002
2879969	198293165	60265	C/G	0.589	0.596	0.8093
958902	198300052	67152	T/C	0.215	0.264	0.0568
1839742	198301232	68332	T/C	0.928	0.946	0.2311
1868890	198304028	71128	T/C	0.420	0.422	0.9494
1868891	198309327	76427	G/A	0.220	0.217	0.8858

[0252] Figure 13 shows the proximal SNPs in and around the DLG1 gene. The position of each SNP on the chromosome is presented on the x-axis. The y-axis gives the negative logarithm (base 10) of the p-value comparing the estimated allele in the case group to that of the control group. The minor allele frequency of the control group for each SNP designated by an X or other symbol on the graphs in Figure 13 can be determined by consulting Table 10. By proceeding down the Table from top to bottom and across the graphs from left to right the allele frequency associated with each symbol shown can be determined.

[0253] To aid the interpretation, multiple lines have been added to the graph. The broken horizontal lines are drawn at two common significance levels, 0.05 and 0.01. The vertical broken lines are drawn every 20kb to assist in the interpretation of distances between SNPs. Two other lines are drawn to

expose linear trends in the association of SNPs to the disease. The light gray line (or generally bottom-most curve) is a nonlinear smoother through the data points on the graph using a local polynomial regression method (W.S. Cleveland, E. Grosse and W.M. Shyu (1992) Local regression models. Chapter 8 of Statistical Models in S eds J.M. Chambers and T.J. Hastie, Wadsworth & Brooks/Cole.). The black line (or generally top-most curve, *e.g.*, see peak in left-most graph just to the left of position 92150000) provides a local test for excess statistical significance to identify regions of association. This was created by use of a 10kb sliding window with 1kb step sizes. Within each window, a chi-square goodness of fit test was applied to compare the proportion of SNPs that were significant at a test wise level of 0.01, to the proportion that would be expected by chance alone (0.05 for the methods used here). Resulting p-values that were less than 10^{-8} were truncated at that value.

[0254] Finally, the gene or genes present in the loci region of the proximal SNPs as annotated by Locus Link (http address: www.ncbi.nlm.nih.gov/LocusLink/) are provided on the graph. The exons and introns of the genes in the covered region are plotted below each graph at the appropriate chromosomal positions. The gene boundary is indicated by the broken horizontal line. The exon positions are shown as thick, unbroken bars. An arrow is place at the 3' end of each gene to show the direction of transcription.

Example 4

KIAA0783 Proximal SNPs

[0255] It has been discovered that a polymorphic variation (rs220097) in a region that encodes KIAA0783 is associated with the occurrence of breast cancer (see Examples 1 and 2). Subsequently, SNPs proximal to the incident SNP (rs220097) were identified and allelotyped in breast cancer sample sets and control sample sets as described in Examples 1 and 2. Approximately fifty-eight allelic variants located within the KIAA0783 region were identified and allelotyped. The polymorphic variants are set forth in Table 11.

Table 11

dbSNP rs#	Chromosome	Position in Figure 2	Chromosome Position	Allele Variants
218973	7	201	10710201	G/A
218962	7	6395	10716395	T/C
1640705	7	8558	10718558	T/C
218983	7	9429	10719429	C/T
190075	7	9809	10719809	T/G
284856	7	10072	10720072	C/T
218981	7	10511	10720511	C/T
218980	7	11556	10721556	C/G
1640703	7	16857	10726857	A/G
1640702	7	16951	10726951	A/G

dbSNP rs#	Chromosome	Position in Figure 2	Chromosome Position	Allele Variants
1640701	7	17027	10727027	C/G
1681305	7	17177	10727177	T/C
1640700	7	17615	10727615	A/C
1640699	7	17950	10727950	C/G
1154923	7	18329	10728329	T/G
1154922	7	18384	10728384	T/C
1154921	7	18561	10728561	G/A
1154920	7	18579	10728579	C/T
2510348	7	18871	10728871	C/G
1681311	7	27152	10737152	C/T
1681312	7	27306	10737306	T/C
1681286	7	28091	10738091	T/C
1640710	7	28661	10738661	A/C
1681284	7	29011	10739011	T/C
2110377	7	29962	10739962	T/G
2110376	7	29969	10739969	T/G
2160059	7	30085	10740085	T/C
1681290	7	31656	10741656	A/G
1681291	7	31685	10741685	A/G
1681292	7	31749	10741749	G/A
220091	7	45389	10755389	T/C
182594	7	45459	10755459	G/C
220090	7	46647	10756647	A/G
220097	7	49860	10759860	T/C
220096	7	53061	10763061	T/C
220095	7	57308	10767308	T/A
3801435	7	61563	10771563	A/G
1681281	7	61660	10771660	A/G
1026903	7	62212	10772212	C/T
220093	7	67090	10777090	T/G
286243	7	67198	10777198	T/C
3801437	7	70071	10780071	A/G
3801438	7	70191	10780191	G/A
2108111	7	74006	10784006	C/T
2353340	7	75600	10785600	A/G
3823875	7	85761	10795761	A/G
2190295	7	90798	10800798	T/G
KIAA0783_SNP1	7	90883	10800883	C/T
2306768	7	91259	10801259	T/A
2353341	7	95416	10805416	C/G
2353342	7	95446	10805446	T/G
2883140	7	96368	10806368	G/T
2353343	7	97050	10807050	T/C
2108114	7	97362	10807362	C/T
1483204	7	97630	10807630	A/C
1483202	7	97989	10807989	T/C
1483201	7	98107	10808107	C/T

dbSNP rs#	Chromosome	Position in Figure 2	Chromosome Position	Allele Variants
KIAA0783_SNP2	7	NOT MAPPED		

Assay for Verifying and Allelotyping SNPs

[0256] The methods used to verify and allelotype the proximal SNPs of Table 11 are the same methods described in Examples 1 and 2 herein. The PCR primers and extend primers used in these assays are provided in Table 12 and Table 13, respectively.

Table 12

dbSNP rs#	Forward PCR primer	Reverse PCR primer
KIAA0783_SNP1	ACGTTGGATGCCCTAACACTACTCCTTGTC	ACGTTGGATGCCAACACTTCTTGGAGTCTG
KIAA0783_SNP2	ACGTTGGATGAGCCACATTCTCAGATACTG	ACGTTGGATGGGAAAAGAAGGAAGAAGAAG
182594	ACGTTGGATGGAGACAGAAAAGTGGTGGAC	ACGTTGGATGCCTTTAAGAAGGCCCTTGTG
190075	ACGTTGGATGCACAAATTCAGTGGCCAAGC	ACGTTGGATGCTTGTTGTGGACACCTACTG
218962	ACGTTGGATGCAGGAGTGAGAAGTTCTTTG	ACGTTGGATGTGCTGATTGGTCTATGGGTG
218973	ACGTTGGATGTCTCACACTGAGGCCTGTAG	ACGTTGGATGTTTGCTGCACCCATCAACTC
218980	ACGTTGGATGCTTCCCTCCTTTTCTCCTTC	ACGTTGGATGCAAGATCCAATCCAGAAGAC
218981	ACGTTGGATGAGATTGCTGCCACTACACAC	ACGTTGGATGCTCTTGGCATTCTTAACTCAG
218983	ACGTTGGATGTCTGCAGTTTCTCTCTCAAC	ACGTTGGATGACCAAATCCAAGATGTAGGG
220090	ACGTTGGATGCAGCAGAAACTTGATGATGG	ACGTTGGATGAGACACTGAGACTCTGGAGG
220091	ACGTTGGATGGTGTATACACAAGGGCCTTC	ACGTTGGATGCTGATTGCTGTTTCTGTTAC
220093	ACGTTGGATGTCCACACTGTGAACAGAGAC	ACGTTGGATGAGTCTAAAAAGGCTGTCAGG
220095	ACGTTGGATGGCAGCTCAATTTTTAGGAACC	ACGTTGGATGCCCTTGTACACTGTTGCATG
220096	ACGTTGGATGTAGATTAATTATTGGTTGGC	ACGTTGGATGGCCACCTCCAAAATTAGATC
220097	ACGTTGGATGTTCCTGGGAATGGATTTCAG	ACGTTGGATGGCAAACGTGCACATTTGCAC
284856	ACGTTGGATGTGCATGACTACACAAAGAAG	ACGTTGGATGGCAAAATCCTACATTGAGGC
286243	ACGTTGGATGATGTCTCTGTTCACAGTGTG	ACGTTGGATGCTGGCAAATAGCAATCTAAAC
220097	ACGTTGGATGTTCCTGGGAATGGATTTCAG	ACGTTGGATGGCAAACGTGCACATTTGCAC
1026903	ACGTTGGATGGTACTGAAACTCTGAGCATTC	ACGTTGGATGCATCTTATCTGTTTACCATAC
1154920	ACGTTGGATGGCTGTATATACGAGTTAATGG	ACGTTGGATGAGTGGAGGTGGAGGTGAGGCT
1154921	ACGTTGGATGAAATGCCAATAGCGCCAAGG	ACGTTGGATGAGTAGAAGAGATAAGCCTGG
1154922	ACGTTGGATGTTTTGCCTCACCAAGATTGC	ACGTTGGATGACAATTTCATTGAGGAGAGG
1154923	ACGTTGGATGGATGGTTGATCACTTGTGTAG	ACGTTGGATGCTTACCTCCTCTCCTCAATG
1483201	ACGTTGGATGGTTGCTAAAGTAGTTTCAGCC	ACGTTGGATGACCAAAGAGCTTGTCCCATC
1483202	ACGTTGGATGGTGCTTTAGAATGTAACACAG	ACGTTGGATGTGGAATTGCACCCTTGCTTG
1483204	ACGTTGGATGTATCTTATCTAGCAGGCAAC	ACGTTGGATGACTAAGATCACAGGCCTGAG
1640699	ACGTTGGATGGGTTGGGTGTATGATAGGAG	ACGTTGGATGAGCATGGCTAATCTGTCTGG
1640700	ACGTTGGATGCTTTATTGACTGCTTTCAATC	ACGTTGGATGAGTGATTACGAGCCTGTACC
1640701	ACGTTGGATGTTAGGTGCATTGATGCTCTG	ACGTTGGATGCTCAGGCACAGAAAAGATTC
1640702	ACGTTGGATGCTGTGGTCTCAGGTCACAAA	ACGTTGGATGATGCACCTAAAACAAGAGTC

dbSNP rs#	Forward PCR primer	Reverse PCR primer
1640703	ACGTTGGATGCATAATTTACCTTCCTGGCC	ACGTTGGATGCAAATTTGTGACCTGAGACC
1640705	ACGTTGGATGACCATCAGAACCAGTATACC	ACGTTGGATGGATGGCCAGAATTGATGTAC
1640710	ACGTTGGATGCCTTTCCGCTGTATCTCTTG	ACGTTGGATGGGTACAAGGAAGATCCTCAG
1681281	ACGTTGGATGATTGAGAAAGCAGCTGCTTG	ACGTTGGATGCCAACCTCCCAAATACATCC
1681284	ACGTTGGATGATAAAATAGGTCTGGGGCTG	ACGTTGGATGGTTTGCTTACTCTGGTACTG
1681286	ACGTTGGATGGAAATGTAACGCAAAGAGGG	ACGTTGGATGGTTGAAACATTGTCTGCTAG
1681290	ACGTTGGATGGTACCATAAAATACAATACC	ACGTTGGATGTGGTCCCCCAGTCATCTTAA
1681291	ACGTTGGATGTAGCAAAACCCTGCCTCTAC	ACGTTGGATGAGGTCAGTGTTCTGGTATTG
1681292	ACGTTGGATGAGGTCAGTGTTCTGGTATTG	ACGTTGGATGAGCCTGGGCAACATAGCAAA
1681305	ACGTTGGATGCAGACAGATGTTTAGCTACC	ACGTTGGATGTGAAGTTGTGGATTCCCAGC
1681311	ACGTTGGATGGCTTGACCAATCATACTTCC	ACGTTGGATGGAAACAAATTGCTCTGAGTCC
1681312	ACGTTGGATGTCTTCAGGGCAGTAGGATTC	ACGTTGGATGCACATGTGTTTAATACAAGG
2108111	ACGTTGGATGAGCCTGTAAATGATAGAGCC	ACGTTGGATGGATGTCACAGTACAGCAATG
2108114	ACGTTGGATGGATAGAAAAGTTAGAGAAATG	ACGTTGGATGAAGGTCACACCACTGCACTC
2110376	ACGTTGGATGCCAGTTTACACTGGATATTTC	ACGTTGGATGTTGACTAGCTGCTAGAAGGG
2110377	ACGTTGGATGCCAGTTTACACTGGATATTTC	ACGTTGGATGTTGACTAGCTGCTAGAAGGG
2160059	ACGTTGGATGTTAAGTACCGGGAAATTCAG	ACGTTGGATGTCATATACCTACGCAGGCTC
2190295	ACGTTGGATGCTTTTAGAAGTAGTAGGGGC	ACGTTGGATGAGACTCCAAGAAGTGTTGGG
2306768	ACGTTGGATGAAAGGTGGTTTTGCCAGCTG	ACGTTGGATGCTCAGTCTCCTGAAGTGCTG
2353340	ACGTTGGATGCCTATCTGCATGTTGCTTAC	ACGTTGGATGGACTCTTGGGAGTACAAATG
2353341	ACGTTGGATGCACAACCAGAATTTGTAAGTC	ACGTTGGATGCACACGCATGCATCATCTAC
2353342	ACGTTGGATGTGGTTTTCAGTCAAAGCTGC	ACGTTGGATGCTGAGATCTTTCTTCCTGAC
2353343	ACGTTGGATGGTTGCAGAGGGAAGCATTTC	ACGTTGGATGCACTTGTGACCAGGTCACTA
2510348	ACGTTGGATGCTATCCCAGGGCTATGTTTG	ACGTTGGATGGAAGTGGAGGATGAGTTGTG
2883140	ACGTTGGATGCAGCACTTACTTGTCATGTAG	ACGTTGGATGCATAACCAATTTGTCTTAAC
3801435	ACGTTGGATGTCAGTATGAAGCAAGCAGCC	ACGTTGGATGATGTCGCTATACTCTGTAGG
3801437	ACGTTGGATGGTAGCTGAGAAGATGCTCAC	ACGTTGGATGATAGCTGTTCCAGTCTCTTG
3801438	ACGTTGGATGATACGGTAAAGGTAGTCTGG	ACGTTGGATGTTACCTGTATTGCCCTCTCG
3823875	ACGTTGGATGCTCAAGAGCCCATCATCATC	ACGTTGGATGGACAGGCTCAGATATTTCAG

Table 13

dbSNP rs#	Extend Primer	Term Mix
KIAA0783_SNP1	ATTCAGCACAAGTTGTCA	ACG
KIAA0783_SNP2	GAAAGACCTAGAAAGAAAA	ACT
182594	CTCTCTCTTTCTCTCACT	ACT
190075	GTCTGGAGATCCGAATTT	ACT
218962	GCACCATCTGATTGGCC	ACT
218973	CCCAACACTATCCCTTC	ACG
218980	ATCCAGAAGACAATATTGCATTTA	ACT
218981	GTATTGCTTTGTTGCCC	ACG
218983	GGTAAAGAGATGAAGTGC	ACG
220090	CCCAGATATCCTCGGAA	ACT
220091	TGTTACTTATTACATTGTCCAA	ACT
220093	TTATATTCACTCTGAAATCCC	ACT

dbSNP rs#	Extend Primer	Term Mix
220095	CACTGTTGCATGAAATGTA	CGT
220096	CCTGCTACAAAGGGACCTCA	ACT
220097	ACAGAGTTTTAAACCTCCTACA	ACT
284856	TACATTGAGGCAGTTTGTGCT	ACG
286243	AGCAATCTAAACATGAGATTGAGC	ACT
220097	ACAGAGTTTTAAACCTCCTACA	ACT
1026903	CTTATCTGTTTACCATACAATCTA	ACG
1154920	CAACACAAAATGCCAATAG	ACG
1154921	TGTGGCTGTATATACGAGTTAA	ACG
1154922	TTGAGGAGAGGAGGTAA	ACT
1154923	CATCAATCTAATCTCATTTCCTAT	ACT
1483201	TGGGTGGTCCTTTTCTGATA	ACG
1483202	TAATCATGTGGAATTTCCAG	ACT
1483204	CAGGCCTGAGCCACTGT	ACT
1640699	CTAATCTGTCTGGTTAATAGAA	ACT
1640700	GCAAAAGCAAAAGTAAGCT	ACT
1640701	AAACAATGGTAATCTAGAGTAAGC	ACT
1640702	TGATTCAATTTCTGTTGACTACT	ACT
1640703	GTGACCTGAGACCACAGATC	ACT
1640705	TCCAAATAAGAAGCCCT	ACT
1640710	CAGTGTAATAAATTATCAGTTCAT	ACT
1681281	TGGAGTTCAATATAAAGATACAC	ACT
1681284	TGTTTTCAGTTTTATTTGCC	ACT
1681286	TTGTCTGCTAGCCATTT	ACT
1681290	AATCAGTGTTTCTTTAAAGGTC	ACT
1681291	CTGGTATTGTATTTTATGGTACT	ACT
1681292	GGGCAACATAGCAAAACCCTG	ACG
1681305	TTCCCAGCCCTACTTAC	ACT
1681311	CTGAGTCCTAAAAAAAGGT	ACG
1681312	TTAATACAAGGAAATTCCAGC	ACT
2108111	AGAATTTGAAGACATAAAAACC	ACG
2108114	GCGACAGAGCAAGACTC	ACG
2110376	GGGTCAGAGAACTCTATTAA	ACT
2110377	AGAGAACTCTATTAAGTAGGTC	ACT
2160059	CTCATGGATCTGTCTTAC	ACT
2190295	GGGGAAAAAAAAGGTCATATTA	ACT
2306768	CTGAAGTGCTGGGATTATGGG	CGT
2353340	TTTTCTGTGCTTTCTTTGT	ACT
2353341	CATCTACTCTCTTTGAAGTT	ACT
2353342	CTTTCTTCCTGACTTACAAATTC	ACT
2353343	GTGTTTTTGTTGACATATCAAT	ACT
2510348	GGAGGATGAGTTGTGTTGACT	ACT
2883140	TTGTCTTAACTACTATAAACTGAA	CGT
3801435	GCTATACTCTGTAGGAGTTTATCT	ACG

dbSNP rs#	Extend Primer	Term Mix
3801437	CAGTCTCTTGATTTTAAGGA	ACT
3801438	CTCGTACTTTTGCCCAC	ACG
3823875	ATTTCAGTGATATAGGAGTCT	ACT

Genetic Analysis of Allelotyping Results

[0257] Allelotyping results are shown for cases and controls in Table 14. The allele frequency for the A2 allele is noted in the fifth and sixth columns for breast cancer pools and control pools, respectively, where "AF" is allele frequency. The allele frequency for the A1 allele can be easily calculated by subtracting the A2 allele frequency from 1 (A1 AF = 1-A2 AF). For example, the SNP rs218973 has the following case and control allele frequencies: case A1 (G) = 0.640; case A2 (A) = 0.360; control A1 (G) = 0.645; and control A2 (A) = 0.355, where the nucleotide is provided in paranthesis. SNPs with blank allele frequencies were untyped.

Table 14

dbSNP rs#	Position in Figure 2	Chromosome Position	A1/A2 Allele	A2 Case AF	A2 Control AF	p-Value
218973	201	10710201	G/A	0.360	0.355	0.8462
218962	6395	10716395	T/C	0.547	0.535	0.6939
1640705	8558	10718558	T/C	0.601	0.568	0.2583
218983	9429	10719429	C/T	0.561	0.558	0.9406
190075	9809	10719809	T/G	0.447	0.428	0.5348
284856	10072	10720072	C/T	0.612	0.585	0.3555
218981	10511	10720511	C/T	0.432	0.363	0.0189
218980	11556	10721556	C/G	0.409	0.471	0.0378
1640703	16857	10726857	A/G	0.841	0.859	0.3809
1640702	16951	10726951	A/G	0.674	0.656	0.5269
1640701	17027	10727027	C/G	0.266	0.270	0.9020
1681305	17177	10727177	T/C	0.422	0.483	0.0406
1640700	17615	10727615	A/C	0.456	0.423	0.2641
1640699	17950	10727950	C/G	0.344	0.370	0.3558
1154923	18329	10728329	T/G	0.885	0.878	0.7144
1154922	18384	10728384	T/C	0.406	0.479	0.0151
1154921	18561	10728561	G/A	0.367	0.365	0.9611
1154920	18579	10728579	C/T	0.284	0.248	0.1803
2510348	18871	10728871	C/G	0.409	0.425	0.5940
1681311	27152	10737152	C/T	0.251	0.279	0.3099
1681312	27306	10737306	T/C	0.303	0.260	0.1171
1681286	28091	10738091	T/C	0.557	0.544	0.6560
1640710	28661	10738661	A/C	0.455	0.515	0.0472
1681284	29011	10739011	T/C	0.418	0.388	0.3124
2110377	29962	10739962	T/G	0.080	0.058	0.1549
2110376	29969	10739969	T/G	0.265	0.313	0.0798
2160059	30085	10740085	T/C	0.066	0.063	0.8793
1681290	31656	10741656	A/G	0.222	0.287	0.0129
1681291	31685	10741685	A/G	0.017	0.042	0.0143
1681292	31749	10741749	G/A	0.335	0.392	0.0458

dbSNP rs#	Position in Figure 2	Chromosome Position	A1/A2 Allele	A2 Case AF	A2 Control AF	p-Value
220091	45389	10755389	T/C	0.245	0.326	0.0034
182594	45459	10755459	G/C	0.238	0.325	0.0017
220090	46647	10756647	A/G	0.332	0.411	0.0066
220097	49860	10759860	T/C	0.258	0.343	0.0025
220096	53061	10763061	T/C	0.240	0.301	0.0214
220095	57308	10767308	T/A	0.618	0.526	0.0026
3801435	61563	10771563	A/G	0.622	0.507	0.0002
1681281	61660	10771660	A/G	0.501	0.433	0.0235
1026903	62212	10772212	C/T	0.855	0.859	0.8503
220093	67090	10777090	T/G	0.564	0.461	0.0009
286243	67198	10777198	T/C	0.591	0.519	0.0170
3801437	70071	10780071	A/G	0.385	0.459	0.0141
3801438	70191	10780191	G/A	0.018	0.022	0.6491
2108111	74006	10784006	C/T	0.360	0.438	0.0090
2353340	75600	10785600	A/G	0.234	0.309	0.0056
3823875	85761	10795761	A/G	0.502	0.409	0.0025
2190295	90798	10800798	T/G	0.319	0.402	0.0045
KIAA0783_SNP1	90883	10800883	C/T	0.309	0.396	0.0030
2306768	91259	10801259	T/A	0.558	0.472	0.0051
2353341	95416	10805416	C/G	0.163	0.248	0.0008
2353342	95446	10805446	T/G	0.118	0.176	0.0068
2883140	96368	10806368	G/T	0.672	0.561	0.0003
2353343	97050	10807050	T/C	0.071	0.075	0.8073
2108114	97362	10807362	C/T	0.433	0.321	0.0003
1483204	97630	10807630	A/C	0.063	0.093	0.0706
1483202	97989	10807989	T/C	0.643	0.567	0.0101
1483201	98107	10808107	C/T	0.688	0.598	0.0022
KIAA0783_SNP2	NOT MAPPED			0.411	0.459	0.1085

[0258] Figure 14 shows the proximal SNPs in and around the KIAA0783 region. The position of each SNP on the chromosome is presented on the x-axis. The y-axis gives the negative logarithm (base 10) of the p-value comparing the estimated allele in the case group to that of the control group. The minor allele frequency of the control group for each SNP designated by an X or other symbol on the graphs in Figure 14 can be determined by consulting Table 14. By proceeding down the Table from top to bottom and across the graphs from left to right the allele frequency associated with each symbol shown can be determined.

[0259] To aid the interpretation, multiple lines have been added to the graph. The broken horizontal lines are drawn at two common significance levels, 0.05 and 0.01. The vertical broken lines are drawn every 20kb to assist in the interpretation of distances between SNPs. Two other lines are drawn to expose linear trends in the association of SNPs to the disease. The light gray line (or generally bottom-most curve) is a nonlinear smoother through the data points on the graph using a local polynomial regression method (W.S. Cleveland, E. Grosse and W.M. Shyu (1992) Local regression models. Chapter 8 of Statistical Models in S eds J.M. Chambers and T.J. Hastie, Wadsworth & Brooks/Cole.). The black line (or generally top-most curve, *e.g.*, see peak in left-most graph just to the left of position 92150000) provides a local test for excess statistical significance to identify regions of association. This was created

by use of a 10kb sliding window with 1kb step sizes. Within each window, a chi-square goodness of fit test was applied to compare the proportion of SNPs that were significant at a test wise level of 0.01, to the proportion that would be expected by chance alone (0.05 for the methods used here). Resulting p-values that were less than 10^{-8} were truncated at that value.

[0260] Finally, the gene or genes present in the loci region of the proximal SNPs as annotated by Locus Link (http address: www.ncbi.nlm.nih.gov/LocusLink/) are provided on the graph. The exons and introns of the genes in the covered region are plotted below each graph at the appropriate chromosomal positions. The gene boundary is indicated by the broken horizontal line. The exon positions are shown as thick, unbroken bars. An arrow is place at the 3' end of each gene to show the direction of transcription.

Example 5

DPF3 Proximal SNPs

[0261] It has been discovered that a polymorphic variation (rs1990440) in a gene encoding DPF3 is associated with the occurrence of breast cancer (see Examples 1 and 2). Subsequently, SNPs proximal to the incident SNP (rs1990440) were identified and allelotyped in breast cancer sample sets and control sample sets as described in Examples 1 and 2. A total of forty allelic variants located within or nearby the DPF3 gene were identified and allelotyped. The polymorphic variants are set forth in Table 15. The chromosome position provided in column four of Table 15 is based on Genome "Build 33" of NCBI's GenBank.

Table 15

dbSNP rs#	Chromosome	Position in Figure 3	Chromosome Position	Allele Variants
2052146	14	160	71227260	A/C
2052145	14	6053	71233153	T/G
740980	14	9719	71236819	A/G
758915	14	10481	71237581	T/C
758914	14	10676	71237776	A/T
2098195	14	17179	71244279	C/G
740979	14	18561	71245661	A/T
740978	14	18658	71245758	G/C
740977	14	18694	71245794	A/G
740976	14	18858	71245958	T/C
2052143	14	24582	71251682	G/A
2052142	14	24683	71251783	G/A
2052141	14	24767	71251867	C/T
758913	14	27402	71254502	A/G
740975	14	28150	71255250	T/G
747987	14	28494	71255594	T/C
1126160	14	32003	71259103	A/C

dbSNP rs#	Chromosome	Position in Figure 3	Chromosome Position	Allele Variants
2332918	14	35588	71262688	C/T
2332919	14	35619	71262719	T/C
1990443	14	35856	71262956	G/A
3937455	14	36254	71263354	G/C
973963	14	37314	71264414	G/A
1990441	14	40033	71267133	T/G
1990440	14	40095	71267195	G/C
2159715	14	42593	71269693	A/C
2109795	14	42799	71269899	A/G
2159714	14	43090	71270190	G/A
1468662	14	46683	71273783	A/G
2215591	14	49774	71276874	A/G
2109794	14	51796	71278896	C/T
2877821	14	52079	71279179	A/T
2191822	14	53857	71280957	T/C
2191821	14	53971	71281071	A/C
1544579	14	55899	71282999	T/C
2215590	14	60682	71287782	G/A
1004552	14	61291	71288391	C/T
1860749	14	72720	71299820	G/A
1860748	14	72752	71299852	A/C
763388	14	85507	71312607	A/G
1035099	14	89751	71316851	T/A

Assay for Verifying and Allelotyping SNPs

[0262] The methods used to verify and allelotype the sixty-three proximal SNPs of Table 15 are the same methods described in Examples 1 and 2 herein. The PCR primers and extend primers used in these assays are provided in Table 16 and Table 17, respectively.

Table 16

dbSNP rs#	Forward PCR primer	Reverse PCR primer
740975	ACGTTGGATGGAAACCAAGATAGGAAATGG	ACGTTGGATGCTCAGTGCCAGAAATACCAG
740976	ACGTTGGATGTCCTGTTTCTAAGCAGGGAG	ACGTTGGATGATCAGGACTACCTGAGCAAC
740977	ACGTTGGATGTCCAGTGAGGCCTCCCTCCAA	ACGTTGGATGCAGCAACCCAAAGCAACACG
740978	ACGTTGGATGTAGCCACGCCATTATTGGAG	ACGTTGGATGCTTCACATCCCTCCTCAAAG
740979	ACGTTGGATGATCCTAACCAGGTCTGATGG	ACGTTGGATGAAGGGCCAAGCAATGCTTTG
740980	ACGTTGGATGGGTAGGGCTGTCTGTTTCAT	ACGTTGGATGATGCCTGCCACATTGGGTAA
747987	ACGTTGGATGAGGTCTGGCACTGCTAAATG	ACGTTGGATGCCTTGTGAACTTCCAACCTG
758913	ACGTTGGATGCCTAGCCAACATCCTTTTCC	ACGTTGGATGAGCAACCAGTCTAGTTTTCG
758914	ACGTTGGATGCCCTTGTTTTAGAGGTTGGG	ACGTTGGATGTGTGATCCAGACATCAGCTC
758915	ACGTTGGATGCAAGAAGGGCATTTCTACCC	ACGTTGGATGCAATGCTGCTGACATCAGAC

dbSNP rs#	Forward PCR primer	Reverse PCR primer
763388	ACGTTGGATGGGGTACTCTTAGCTGAGAAC	ACGTTGGATGTACAGGGATTGTGATGTGGG
973963	ACGTTGGATGGATTTGTTCTGGCAGGAATG	ACGTTGGATGACAAACCACTAAACTTTCAG
1004552	ACGTTGGATGGATCATCCAAGTATGCTCCC	ACGTTGGATGGCAAAACCCAGTGCCAAAAC
1035099	ACGTTGGATGAAAGGGTACCCAGACTTCAC	ACGTTGGATGTGGGGAGAACTTTGGTCAAC
1126160	ACGTTGGATGGGGTTCTCTCTTGACAGATG	ACGTTGGATGTGTTCTCACCCTGTTCTGTT
1468662	ACGTTGGATGGCTAGAAATCACCAGCAACC	ACGTTGGATGTCATGTAGGTTGGCTCTGAC
1544579	ACGTTGGATGACCATTATCATCTTCCCAGG	ACGTTGGATGCCTTATCTCTCTAAGACATGC
1860748	ACGTTGGATGACTCGACTAGCTAGTCTTGG	ACGTTGGATGAAAGCAATCCAGCGGACAAG
1860749	ACGTTGGATGTCCCCGGAATGATACATGAC	ACGTTGGATGAACATGATTAAGGATAAAGC
1990440	ACGTTGGATGAAGTCACTAACCCCACACAC	ACGTTGGATGCCAGGGTGTGTTCTAATACG
1990441	ACGTTGGATGTCAGAGATATGCACTGCAAG	ACGTTGGATGCACACCCTGGCATGAATGTG
1990443	ACGTTGGATGCACTGGATTTGGCAAGAAGG	ACGTTGGATGTACATGATCCTCCCCTCTAC
2052141	ACGTTGGATGCCTGCAAAATCCCTCATACC	ACGTTGGATGATAGAAGCGTGACCTTACCC
2052142	ACGTTGGATGGGTATGAGGGATTTTGCAGG	ACGTTGGATGACTGGACTCACCCACATAAG
2052143	ACGTTGGATGCCAGTGTAATCACAAGGGTC	ACGTTGGATGTGTGTCACTTCTACCTCCAC
2052145	ACGTTGGATGGTGCTGGCTGCCTAGTTCTA	ACGTTGGATGGGCTTCTCAATTCAGATGGG
2052146	ACGTTGGATGCCACAAAAGCACGTGATTTC	ACGTTGGATGTTATTTGAGCTCTGATAGTG
2098195	ACGTTGGATGGCTCCAGTCTCTAATCACAC	ACGTTGGATGCAAAGTTCTCTGCCTGAGTG
2109794	ACGTTGGATGTAATCCCAGCACTTTGGGAG	ACGTTGGATGAGGCTGATCTTGAACTCCTG
2109795	ACGTTGGATGCAAACAAGGTCCCAGCATTC	ACGTTGGATGTCCTGACTCTCTCAAAACCC
2159714	ACGTTGGATGAAACTCTCTCGTTGCTGTGG	ACGTTGGATGAAAGCCCCTCTAGCAAAAGG
2159715	ACGTTGGATGCTGCCTGCAAGTTCCCATTG	ACGTTGGATGTACAGGCACTGGCGAAGAAG
2191821	ACGTTGGATGGAAAGTGTCCTTAGCTTGCC	ACGTTGGATGTGAGATGGATCTGGAGCCAC
2191822	ACGTTGGATGATTTTTCCCGGCATCTGACC	ACGTTGGATGTGCAAAGTGGTGGAGGAAAG
2215590	ACGTTGGATGTCCAAGAAGGACAGCAGTAG	ACGTTGGATGATGAGAGCCTTTCTTCAGGG
2215591	ACGTTGGATGATTTGTTAAAATTCATAGAAC	ACGTTGGATGTCCCCAGTTTGCATCTTGAC
2332918	ACGTTGGATGAACCCATGGGACCACAATTC	ACGTTGGATGTAGGATGGGTGTTTCCTAGC
2332919	ACGTTGGATGTCTGAGGGCTCTCTCTAATG	ACGTTGGATGATGAAGGAAGAAGCCCTGAC
2877821	ACGTTGGATGATAATCTATGTCCTAGATTG	ACGTTGGATGTAGTAGCATTCCAAGTGCCC
3937455	ACGTTGGATGGCAAGAATAGGTTCTTTCGC	ACGTTGGATGACCTCCACACTCATTACCTC

Table 17

dbSNP rs#	Extend Primer	Term Mix
740975	ACCAGCTCTCTTTGGAT	ACT
740976	ATCCAGATGGCCCTGAC	ACT
740977	TGGTTTTCGAATAAGTAGCCAC	ACT
740978	AAGCCTTCCTATCCCCA	ACT
740979	TGCTTTGGGGCAGACTGAC	CGT
740980	CACATTGGGTAAATGATGA	ACT
747987	AACCTGGTTCTGCCATT	ACT
758913	CCAGTCTAGTTTTCGATCACC	ACT
758914	CCCCAGTGATCCTGAGAAAT	CGT

dbSNP rs#	Extend Primer	Term Mix
758915	GACATCAGACCTATGCCAGGA	ACT
763388	CACTCATGCCTCAAGCCAAT	ACT
973963	AACAACCAACTCTCCAG	ACG
1004552	TCTTGGCTCAGTGCTGC	ACG
1035099	TTGGTCAACATCGCAGC	CGT
1126160	GAAGCCCATCGCTAAGTGTTT	ACT
1468662	CTCTGACTGAGGAGAGACC	ACT
1544579	GACATGCATCAAAGCAGCTG	ACT
1860748	TCTTGGAGCCATATTTTATTTG	ACT
1860749	TTAAGGATAAAGCAATCCAG	ACG
1990440	CGTCAGCAAATGTGTACCGA	ACT
1990441	CATGAATGTGATTCACATTCTCC	ACT
1990443	TTCCCCTCAGCTCTTAG	ACG
2052141	CTTACCCCCAAAGATGTCCA	ACG
2052142	AGCCAGGATAATCTCCTCA	ACG
2052143	TCTACCTCCACTTCCAA	ACG
2052145	ATTCAGATGGGATCACAGAAG	ACT
2052146	GAGCTCTGATAGTGATTGTGAGT	ACT
2098195	TAAACCTTTCTATGTTCCTG	ACT
2109794	CTCAGGTGATCCACCCA	ACG
2109795	TCCCAGAATTTGGAGCC	ACT
2159714	CAAAAGGATCTGCAAAAG	ACG
2159715	CATAGGGATAGGAATGGG	ACT
2191821	ATGTGGGTTTGGACTGGGGCT	ACT
2191822	AGGAAAGGAATGTCTGCCCC	ACT
2215590	CAGGGCCAGCCATGAACGT	ACG
2215591	TTCAATAAAATGTACTCATTCAAA	ACT
2332918	TCTCTCTAATGGGGACC	ACG
2332919	ACTGGATCCCAGAAGAG	ACT
2877821	CCCTGTTCTGCACCTTTAAA	CGT
3937455	TCCTTTTTTCCCCACCC	ACT

Genetic Analysis of Allelotyping Results

[0263] Allelotyping results are shown for cases and controls in Table 18. The allele frequency for the A2 allele is noted in the fifth and sixth columns for breast cancer pools and control pools, respectively, where "AF" is allele frequency. The allele frequency for the A1 allele can be easily calculated by subtracting the A2 allele frequency from 1 (A1 AF = 1-A2 AF). For example, the SNP in row 2 of Table 13 (rs2052146) has the following case and control allele frequencies: case A1 (A) = 0.990; case A2 (C) = 0.010; control A1 (A) = 0.948; and control A2 (C) = 0.052, where the nucleotide is provided in parenthesis. SNPs with blank allele frequencies were untyped ("not AT").

Table 18

dbSNP rs#	Position in Fig 3	Chr m Position	Alleles (A1/A2)	A2 Case AF	A2 Contr l AF	p-Value
2052146	160	71227260	A/C	0.010	0.042	0.0014
2052145	6053	71233153	T/G	0.858	0.776	0.0007
740980	9719	71236819	A/G	0.620	0.644	0.4134
758915	10481	71237581	T/C	0.718	0.718	0.9903
758914	10676	71237776	A/T	0.754	0.749	0.8560
2098195	17179	71244279	C/G	0.976	0.989	0.1034
740979	18561	71245661	A/T	0.656	0.694	0.1850
740978	18658	71245758	G/C	0.011	0.047	0.0005
740977	18694	71245794	A/G	0.913	0.873	0.0353
740976	18858	71245958	T/C	0.610	0.676	0.0217
2052143	24582	71251682	G/A	0.466	0.405	0.0418
2052142	24683	71251783	G/A	0.015	0.051	0.0011
2052141	24767	71251867	C/T	0.363	0.315	0.0950
758913	27402	71254502	A/G	0.931	0.871	0.0011
740975	28150	71255250	T/G	0.461	0.514	0.0763
747987	28494	71255594	T/C	0.715	0.813	0.0003
1126160	32003	71259103	A/C	0.349	0.409	0.0392
2332918	35588	71262688	C/T	0.041	0.070	0.0355
2332919	35619	71262719	T/C	0.300	0.271	0.2797
1990443	35856	71262956	G/A	0.324	0.268	0.0407
3937455	36254	71263354	G/C	0.445	0.455	0.7518
973963	37314	71264414	G/A	0.029	0.035	0.6030
1990441	40033	71267133	T/G	0.128	0.152	0.2380
1990440	40095	71267195	G/C	0.744	0.842	0.0002
2159715	42593	71269693	A/C	0.534	0.542	0.7822
2109795	42799	71269899	A/G	0.795	0.747	0.0582
2159714	43090	71270190	G/A	0.035	0.036	0.9187
1468662	46683	71273783	A/G	0.035	0.069	0.0118
2215591	49774	71276874	A/G	0.892	0.857	0.0776
2109794	51796	71278896	C/T	0.042	0.041	0.9714
2877821	52079	71279179	A/T	0.778	0.862	0.0005
2191822	53857	71280957	T/C	0.899	0.845	0.0078
2191821	53971	71281071	A/C	0.427	0.422	0.8733
1544579	55899	71282999	T/C	0.496	0.483	0.6724
2215590	60682	71287782	G/A	0.271	0.285	0.5936
1004552	61291	71288391	C/T	0.393	0.378	0.5996
1860749	72720	71299820	G/A	0.652	0.522	0.0001
1860748	72752	71299852	A/C	0.894	0.820	0.0007
763388	85507	71312607	A/G	0.291	0.310	0.4883
1035099	89751	71316851	T/A	0.555	0.543	0.7079

[0264] Figure 15 shows the proximal SNPs in and around the DPF3 gene. As indicated, some of the SNPs were untyped. The position of each SNP on the chromosome is presented on the x-axis. The y-axis gives the negative logarithm (base 10) of the p-value comparing the estimated allele in the case group to that of the control group. The minor allele frequency of the control group for each SNP designated by an X or other symbol on the graphs in Figure 15 can be determined by consulting Table 18. By proceeding down the Table from top to bottom and across the graphs from left to right the allele frequency associated with each symbol shown can be determined.

[0265] To aid the interpretation, multiple lines have been added to the graph. The broken horizontal lines are drawn at two common significance levels, 0.05 and 0.01. The vertical broken lines are drawn every 20kb to assist in the interpretation of distances between SNPs. Two other lines are drawn to expose linear trends in the association of SNPs to the disease. The light gray line (or generally bottom-most curve) is a nonlinear smoother through the data points on the graph using a local polynomial regression method (W.S. Cleveland, E. Grosse and W.M. Shyu (1992) Local regression models. Chapter 8 of Statistical Models in S eds J.M. Chambers and T.J. Hastie, Wadsworth & Brooks/Cole.). The black line (or generally top-most curve, *e.g.*, see peak in left-most graph just to the left of position 92150000) provides a local test for excess statistical significance to identify regions of association. This was created by use of a 10kb sliding window with 1kb step sizes. Within each window, a chi-square goodness of fit test was applied to compare the proportion of SNPs that were significant at a test wise level of 0.01, to the proportion that would be expected by chance alone (0.05 for the methods used here). Resulting p-values that were less than 10^{-8} were truncated at that value.

[0266] Finally, the gene or genes present in the loci region of the proximal SNPs as annotated by Locus Link (http address: www.ncbi.nlm.nih.gov/LocusLink/) are provided on the graph. The exons and introns of the genes in the covered region are plotted below each graph at the appropriate chromosomal positions. The gene boundary is indicated by the broken horizontal line. The exon positions are shown as thick, unbroken bars. An arrow is place at the 3' end of each gene to show the direction of transcription.

Example 6

CENCP1 Proximal SNPs

[0267] It has been discovered that a polymorphic variation (rs355510) in the CENPC1 region is associated with the occurrence of breast cancer (see Examples 1 and 2). Subsequently, SNPs proximal to the incident SNP (rs355510) were identified and allelotyped in breast cancer sample sets and control sample sets as described in Examples 1 and 2. Approximately seventy-nine allelic variants located within the CENPC1 region were identified and allelotyped. The polymorphic variants are set forth in Table 19. The chromosome position provided in column four of Table 19 is based on Genome "Build 33" of NCBI's GenBank.

Table 19

dbSNP rs#	Chromosome	Position in Figure 4	Chromosome Position	Allele Variants
1874633	4	196	68275196	A/G
1846060	4	13311	68288311	G/A
451352	4	14486	68289486	C/T
355468	4	14691	68289691	A/T
355469	4	15551	68290551	C/G

dbSNP rs#	Chromosome	Position in Figure 4	Chromosome Position	Allele Variants
355470	4	17702	68292702	T/C
355471	4	17872	68292872	T/C
191650	4	19588	68294588	T/C
355472	4	19910	68294910	T/A
1874635	4	20006	68295006	A/C
1497430	4	20575	68295575	A/G
2254659	4	21092	68296092	G/A
3822197	4	22830	68297830	C/T
2632453	4	23455	68298455	A/G
2646282	4	23716	68298716	G/A
2646285	4	23890	68298890	T/G
768244	4	24001	68299001	C/T
724199	4	24995	68299995	G/A
1187960	4	27282	68302282	T/C
1187961	4	27779	68302779	C/T
355518	4	29099	68304099	C/G
355519	4	31185	68306185	A/G
355511	4	33994	68308994	C/T
451397	4	34942	68309942	T/C
355513	4	35137	68310137	C/G
355514	4	36538	68311538	T/C
355515	4	37139	68312139	C/T
1056789	4	37358	68312358	G/A
2646290	4	38828	68313828	A/G
190255	4	39469	68314469	T/C
355466	4	40233	68315233	T/C
355465	4	40472	68315472	A/T
2646292	4	41679	68316679	C/T
2632454	4	41682	68316682	G/A
1056787	4	42831	68317831	A/G
CENPC1_SNP1	4	42976	68317976	A/G
173317	4	44128	68319128	A/G
451344	4	44195	68319195	C/T
355510	4	46769	68321769	G/A
355508	4	47363	68322363	G/C
451391	4	48843	68323843	C/T
355500	4	52574	68327574	A/G
355499	4	52602	68327602	A/G
355498	4	53212	68328212	A/G
1187974	4	53781	68328781	C/G
355493	4	54710	68329710	A/T
2632456	4	55808	68330808	G/A
1825790	4	57987	68332987	T/A
355475	4	58556	68333556	C/A
1391110	4	59148	68334148	T/A
1442557	4	59286	68334286	G/C
355478	4	60217	68335217	A/G

dbSNP rs#	Chromosome	Position in Figure 4	Chromosome Position	Allele Variants
189579	4	60412	68335412	G/T
355480	4	60753	68335753	C/T
355481	4	60791	68335791	T/G
355483	4	61524	68336524	A/G
355485	4	62543	68337543	T/C
2646267	4	62825	68337825	A/G
2646268	4	62826	68337826	A/C
355486	4	62857	68337857	C/T
355487	4	63400	68338400	T/C
355488	4	63960	68338960	T/A
355489	4	64307	68339307	A/G
451376	4	64539	68339539	A/G
1353626	4	65728	68340728	A/G
2632450	4	66000	68341000	G/A
2646269	4	66521	68341521	T/G
2276945	4	68185	68343185	C/T
3775861	4	69643	68344643	G/A
1403151	4	74909	68349909	C/A
1843833	4	82973	68357973	T/G
1843831	4	83039	68358039	T/C
3806810	4	85713	68360713	A/G
3775862	4	86873	68361873	T/C
1962700	4	90293	68365293	T/G
2046601	4	91810	68366810	T/G
2171386	4	92609	68367609	A/G
2046599	4	92884	68367884	G/A
355490	4			A/T

Assay for Verifying and Allelotyping SNPs

[0268] The methods used to verify and allelotype the proximal SNPs of Table 19 are the same methods described in Examples 1 and 2 herein. The PCR primers and extend primers used in these assays are provided in Table 20 and Table 21, respectively.

Table 20

dbSNP rs#	Forward PCR primer	Reverse PCR primer
1056787	ACGTTGGATGCATTTCATATTTTGTAGATC	ACGTTGGATGTCTCAGCCCTCTGATAAAAC
1056789	ACGTTGGATGTGAAGGTTCTGGAGGTATCG	ACGTTGGATGTCTTCTTAGCCAAGTCTGCC
CENPC1_SNP1	ACGTTGGATGAACAACGCACAATATCCCCG	ACGTTGGATGGGGTGAGGTTTATGGGAATG
11250	ACGTTGGATGAACAACGCACAATATCCCCG	ACGTTGGATGCATTTGCCAAAGTCTTAGGT
1187960	ACGTTGGATGTGAACCCTTCAAAATCACCC	ACGTTGGATGTTGTGTTTCATGGGAGGAGG
1187961	ACGTTGGATGCAACAGATTTTCCCTGTAGAC	ACGTTGGATGTGCATTGACTTCTCCTCAGC
1187974	ACGTTGGATGGCTGAGCAGAAGCTCTTTCA	ACGTTGGATGTGGGCAAAGACTTCATGATT

dbSNP rs#	Forward PCR primer	Reverse PCR primer
1353626	ACGTTGGATGCAACTACTACCTAGATGATGA	ACGTTGGATGAATAGAAAATCTAAATTGTCTAC
1391110	ACGTTGGATGAGTATGAAGGTCAGGGTCAG	ACGTTGGATGAAAGAGCACTGACCATGGAG
1403151	ACGTTGGATGTCAGTCAGAGATCATAGTTC	ACGTTGGATGCATGTAGTGCTTTAACAAATG
1442557	ACGTTGGATGCAACACATGCACCATTAGCG	ACGTTGGATGGAAGCCACAAACAGATCAGG
1497430	ACGTTGGATGTTGCTTGCTTGATGATTGGC	ACGTTGGATGTCTTCTGGACTTTAGCACTG
173317	ACGTTGGATGCTATAGGACTGTAAATTGTAG	ACGTTGGATGTTTTTACACACATGCTGTCA
1825790	ACGTTGGATGGGCCAACATGGTAAAACTCC	ACGTTGGATGCTGGGATAACAGGTACTTGC
1843831	ACGTTGGATGTCTCAGCTCATTTCCACCTC	ACGTTGGATGACCTGTAGTCCCAGCTACTC
1843833	ACGTTGGATGGACCAACATGGTGAAATCTC	ACGTTGGATGTGAGTAGCTGGGACTACAGG
1846060	ACGTTGGATGAAGATTATCACCGCACTGGG	ACGTTGGATGATCTCCTGACCTCGTGATCC
1874633	ACGTTGGATGAGGTTTTTGGTATGGTTAGC	ACGTTGGATGGAAAAGGGAGTTGGCCTAAA
1874635	ACGTTGGATGAGAGAGAGAGAGAGAGAGAG	ACGTTGGATGATGGGCTATAGTGGGATAGG
189579	ACGTTGGATGACACCAAAAGCAATGGCAAC	ACGTTGGATGGTTGCCTGTTCACTCTGATG
190255	ACGTTGGATGGAGATCTAGCACATTTATCC	ACGTTGGATGAGGTTGCCTGAAATGCTAAG
191650	ACGTTGGATGGAGATACCTTTGCTAAGGTG	ACGTTGGATGGGTAGTAATAATGGTACTCC
1962700	ACGTTGGATGATAAGAGAGAGTGTGGGTGG	ACGTTGGATGATTTCCTGACCTCGTGATCC
2046599	ACGTTGGATGTATTGAATTCCCTCTGTATG	ACGTTGGATGTCATTCTTTTGAGACTGAAC
2046601	ACGTTGGATGGCTCCAATGACTAAGTGGAC	ACGTTGGATGGACAGAACACTAAGAGCCTA
2171386	ACGTTGGATGCTTATCGAAATGAAATCAAG	ACGTTGGATGACAGCTGCAAACCTAAGGAC
2254659	ACGTTGGATGATCTCTAAGTGAGATAGAGG	ACGTTGGATGCCCAGTCAAATGAAACCCAC
2276945	ACGTTGGATGGGGAATTCTATATTCCCATTG	ACGTTGGATGCCCAATTCCAACAGAAAATATC
2632450	ACGTTGGATGTTGAGACAAGCCTAGGCAAC	ACGTTGGATGGTGCTGGGATTACAGGTGTG
2632453	ACGTTGGATGAAAAGTGAGAGGGCAATAGG	ACGTTGGATGCATAGTAAGTCACCACAAGC
2632454	ACGTTGGATGTTCTGTGGGTCAGATGTCTC	ACGTTGGATGAGAAACAGACTTCCTCCCAG
2632456	ACGTTGGATGCCACCATATCAACAGATCAG	ACGTTGGATGCCTGCCAGTATGCTGAGAAT
2646267	ACGTTGGATGTGAGAAAAAGCACTCCTGGG	ACGTTGGATGAGGCTGAGACAGGAGAATTG
2646268	ACGTTGGATGCAGGAGAATTGCTTGAACCC	ACGTTGGATGTGAGAAAAAGCACTCCTGGG
2646269	ACGTTGGATGACCACTATTGTTTCTTTCTC	ACGTTGGATGGGCTAAAGAGTGAAACCCTG
2646282	ACGTTGGATGGATTGTTTTGAGTCATCTAC	ACGTTGGATGCTGAAATTGACCAGGAAACAC
2646285	ACGTTGGATGGGTGGATTGGACAAACTTGC	ACGTTGGATGCCTTTTGCTTTTCATTGCTC
2646290	ACGTTGGATGGATAGCAAGCTACCTAAGAC	ACGTTGGATGCCTCCTTACTCCACTCAATC
2646292	ACGTTGGATGTTCTGTGGGTCAGATGTCTC	ACGTTGGATGCAAAGAAACAGACTTCCTCC
355465	ACGTTGGATGTATGAGGTTCTGCCACCAAG	ACGTTGGATGTACCAAATCTGAGGGTAGTC
355466	ACGTTGGATGCAGGAGCTGCTTAATTCCTC	ACGTTGGATGGATCTTGGGCACTAAGTCTC
355468	ACGTTGGATGCCTCTCCTCATTTCTGTAAAC	ACGTTGGATGGGCAGGTGGTTAGCATTAAG
355469	ACGTTGGATGTTGGGATCTAGGCATCAAGG	ACGTTGGATGAGGAGGCACATAATGCTTGG
355470	ACGTTGGATGACATACACACACACACACAC	ACGTTGGATGGAGACATACACCTCTGCAAC
355471	ACGTTGGATGCTCATTACAACTTCAGCCAG	ACGTTGGATGACTCAGGACTAAGCTAGTTG
355472	ACGTTGGATGTCTCTCTCTCTCTCTCTCTC	ACGTTGGATGCAGCCCTTAGTACTCAATGG
355475	ACGTTGGATGCTGTCTTATCCCAACTTAGA	ACGTTGGATGGTCATGTTACATACCGAAAC
355478	ACGTTGGATGGGAGGAATCCATATATAGGC	ACGTTGGATGCTGCTGAAGGGAATGAGTAC
355480	ACGTTGGATGGTTTACAGTCCCACCAACAG	ACGTTGGATGAGTCAGGAAACAACAGGTGC
355481	ACGTTGGATGATTGCCACACTGTCTTCCAC	ACGTTGGATGGGATGTGGAGAAACAGGAAC
355483	ACGTTGGATGCCATGTAAGTCTGTCATTTA	ACGTTGGATGAAGTGGTAGCAGAAGTGTGG
355485	ACGTTGGATGAAGAAGAGGCATGCAAACAG	ACGTTGGATGCTGCGACAAAAGACACATTC
355486	ACGTTGGATGTGAGAAAAAGCACTCCTGGG	ACGTTGGATGAGGAGAATTGCTTGAACCCG
355487	ACGTTGGATGCGAGGTAATGAGCAAAGTAAG	ACGTTGGATGGACATTAGGTTCATCTAACCC

dbSNP rs#	Forward PCR primer	Reverse PCR primer
355488	ACGTTGGATGCCAGTTTTCTATGACAAACG	ACGTTGGATGAAAGAGCAGGGACAGCAAAG
355489	ACGTTGGATGACTCTAGGTATTTTGACTCC	ACGTTGGATGAACTTCCATAGTAGAAAGCC
355490	ACGTTGGATGAACTTCCATAGTAGAAAGCC	ACGTTGGATGACTCTAGGTATTTTGACTCC
355493	ACGTTGGATGAGTGGTTTGCTGCACCTATC	ACGTTGGATGGGGAGAGCATTAGGACAAAC
355498	ACGTTGGATGATGAGAGAGGACACAAAGAG	ACGTTGGATGTTACTTTGCACAGTGTGGCC
355499	ACGTTGGATGCAATCAAGCAGAAGGATGGG	ACGTTGGATGGGTGTCTTCTTATAGTTGTC
355500	ACGTTGGATGCAATCAAGCAGAAGGATGGG	ACGTTGGATGGGTGTCTTCTTATAGTTGTC
355508	ACGTTGGATGGTGTAGATGTGTATCAGGTCA	ACGTTGGATGGTCCACAAAGCATAGCATCC
355510	ACGTTGGATGCCCTCCTTTTAACCTTTTAGG	ACGTTGGATGTTCTGAGATGATCCTGATGG
355511	ACGTTGGATGCAGGAGGATATGTGAAAGTC	ACGTTGGATGGTGGATACCAAAATCCAAGG
355513	ACGTTGGATGTGCTGTATAACAGATTACCC	ACGTTGGATGAACTAGCTAGCTAAGCCTCC
355514	ACGTTGGATGCCTCAATAGGTTGTTGGAAC	ACGTTGGATGTTGAGTTCATACTATGTGCC
355515	ACGTTGGATGAGCTCTGCACTCTGACATAC	ACGTTGGATGGTGCAGAGTACTACTTTGCC
355518	ACGTTGGATGTGCCATGGGGTTGTAAAATC	ACGTTGGATGACACAGAGACCAGCTGAAAG
355519	ACGTTGGATGGGGAAGAAGCAGATTTTGAG	ACGTTGGATGCATAGGTTGAGAACATCAAGC
3775861	ACGTTGGATGCCATCTCTTTGAAAATTCCAC	ACGTTGGATGCCCTCAAGTACTTGTTTTGTC
3775862	ACGTTGGATGTAATGAAGCTGAGTTTATTC	ACGTTGGATGGTTTTTGTTTATTGGTGTCC
3806810	ACGTTGGATGTCTTTTCTCCCATCATTTCC	ACGTTGGATGACTCAATGGTTGCATGTAGG
3822197	ACGTTGGATGTGTTTGCTAAAGCTATGCTG	ACGTTGGATGTGAGCATTATGCCTAAGAGC
451344	ACGTTGGATGCCTTTCTAGATACACTCCAT	ACGTTGGATGCAGCATGTGTGTAAAAATGC
451352	ACGTTGGATGAGGCAAATTATTTTTGGATG	ACGTTGGATGCTCCCTAAATGGGGAAAAAAG
451362	ACGTTGGATGCAACACATGCACCATTAGCG	ACGTTGGATGGAAGCCACAAACAGATCAGG
451376	ACGTTGGATGAGCAGTCTATTCTGGTTCAC	ACGTTGGATGGCCTTTGAGCTTTAAAAATC
451391	ACGTTGGATGTAAAGTAGGGACTGGGATGG	ACGTTGGATGGCTGTAGAGTAGTGAAACCC
451397	ACGTTGGATGGTTGCCATATTCAGCAGCTG	ACGTTGGATGCTGTTTCCAGTAGACCTTAG
724199	ACGTTGGATGCCAGCTAAAACTGCAAATAC	ACGTTGGATGTGGACTCATTTGAGAATATG
768244	ACGTTGGATGTAAAACCCCTTCCTCATCCC	ACGTTGGATGACCTTTAGCAGCCTGAAACC

Table 21

dbSNP rs#	Extend Primer	Term Mix
355469	GCACATAATGCTTGGTTGTATT	ACT
CENPC1_SNP1	CTTGACTTTCTACCTTGAA	ACT
11250	CTCTTGACTTTCTACCTTGAA	ACT
173317	ACTTAGCGGCTTAAAACAAC	ACT
189579	CTGTTCACTCTGATGGTAGTTT	CGT
190255	GTACTATGTGGCAGATGA	ACT
191650	GGTACTCCTACTTAAAATTTTG	ACT
355465	GAGGGTAGTCTTGGGAACC	CGT
355466	CTCTAGTGAGCTTCCCT	ACT
355468	AGCATTAAGTATTCATGAGAGTTC	CGT
355470	GGTCTGTTTTATATGTGTGT	ACT
355471	AGCTAGTTGCTTCAGTAAGT	ACT
355472	GTACAGTCATAACAGTTGTTAA	CGT

dbSNP rs#	Extend Primer	Term Mix
355475	TACATACCGAAACACATTCC	CGT
355478	ACATTCTATATGGCCCCTTG	ACT
355480	GGAGAGGATGTGGAGAAA	ACG
355481	GGTGGGACTGTAAACTA	ACT
355483	AGAAGTGTGGACACAGTATC	ACT
355485	CACATTCAACTATACACGCTTTTA	ACT
355486	GTGAGCCGAAATCGTGCCAC	ACG
355487	TTCATCTAACCCTTTTCATAA	ACT
355488	AGCAAAGCTGAAAATGATAA	CGT
355489	CAATAAATAATAGCAAAGACTGG	ACT
355490	TGTTTATATTGCTGTTTCTTGA	CGT
355493	CTCATGTGGGGCTTAAA	CGT
355498	GTGTGGCCATTTTCACT	ACT
355499	TGTTAGATAGAGGTTTATCATTTT	ACT
355500	TTTTTCCTGCAATAGTTTTCT	ACT
355508	ATACTTATGCTCTGCTACC	ACT
355510	ATGGTTTTCTTTCTTGTCCTTC	ACG
355511	GGATGCTCAAGTCCCTTATATA	ACG
355513	GCCTCCCAGATTGCTGA	ACT
355514	TGTGCCAAATATTTGCTAGAT	ACT
355515	ACTACTTTGCCTGTGTGTCA	ACG
355518	ACCAGCTGAAAGAAAATC	ACT
355519	AAGCTTAGTATGTCCAAATCTAAC	ACT
451344	GTGTGTAAAAATGCATTCCAAGTT	ACG
451352	CCCCCGAAATGTTTCAAAGG	ACG
451362	CCACAAACAGATCAGGTTGGTG	ACT
451376	AGTATGTAAAAAGATAGGGAAGA	ACT
451391	GAGTAGTGAAACCCCTGACC	ACG
451397	CAGTAGACCTTAGTTTCTTAACC	ACT
724199	GAGAATATGATAAAAGCTCAGACC	ACG
768244	GTTTCTGTCTCTGGCGA	ACG
1056787	GGATACAAGTTATGCTTTGATAG	ACT
1056789	TCCAATGGCTCACTCAG	ACG
1187960	GGAGGAGGTCAAAATATCA	ACT
1187961	GACTTCTCCTCAGCTATGAA	ACG
1187974	TGATTAAAACACCAAAAGCAATT	ACT
1353626	AATCTAAATTGTCTACTGAAACT	ACT
1391110	CCATGGAGTTGTAAGGAA	CGT
1403151	TAGTGCTTTAACAAATGCTGTCA	CGT
1442557	CACAAACAGATCAGGTTGGTG	ACT
1497430	GAATTGGGGAGAGAAAGGGA	ACT
1825790	CCTGGCAAATTTTGGTATTTTTAG	CGT
1843831	GCGGGAGAATGGCATGA	ACT
1843833	GCTCACCACCACACCTG	ACT

dbSNP rs#	Extend Primer	Term Mix
1846060	AAAGTGCTGGGATTACAGG	ACG
1874633	TGGCCTAAAAATATTTTTACCGT	ACT
1874635	CAACTGTTTAACAACCAGGC	ACT
1962700	AGAGTGCTGGGATTACA	ACT
2046599	CTTTTGAGACTGAACACCTCTA	ACG
2046601	AGAACACTAAGAGCCTAGAATGG	ACT
2171386	AGTATGCAGAGACTTACAG	ACT
2254659	AACCCACCATTCCTATG	ACG
2276945	CACAAAATACCTCCAAATTTTA	ACG
2632450	TTACAGGTGTGAGCCAC	ACG
2632453	CACCACAAGCCACTTGA	ACT
2632454	CTTCCTCCCAGAGCCAC	ACG
2632456	TCATAGGTAATGTGGATTTTGT	ACG
2646267	TTGCTTGAACCCGGGAG	ACT
2646268	TCGGCTCACTGCAATCTCT	ACT
2646269	TTCTCGCAAAGAGAAAAC	ACT
2646282	GGAATTAGCAGTCATTTCTTA	ACG
2646285	ATTTCTCTAGACTTTGCTACAAT	ACT
2646290	AGTTCATCCTTCAGGAA	ACT
2646292	AGACTTCCTCCCAGAGC	ACG
3775861	GTTTTGTCTTCAAATAGTAAAGA	ACG
3775862	TCCATTTTTATTTGCAGAAGAC	ACT
3806810	ATTGGATTTGGCGTAGC	ACT
3822197	AGCAGTAGGCAACTTCT	ACG

Genetic Analysis of Allelotyping Results

[0269] Allelotyping results are shown for cases and controls in Table 22. The allele frequency for the A2 allele is noted in the fifth and sixth columns for breast cancer pools and control pools, respectively, where "AF" is allele frequency. The allele frequency for the A1 allele can be easily calculated by subtracting the A2 allele frequency from 1 (A1 AF = 1-A2 AF). For example, the SNP rs1874633 has the following case and control allele frequencies: case A1 (A) = 0.514; case A2 (G) = 0.486; control A1 (A) = 0.449; and control A2 (G) = 0.551, where the nucleotide is provided in paranthesis. SNPs with blank allele frequencies were untyped.

Table 22

dbSNP rs#	Position in Figure 4	Chromosome Position	A1/A2 Allele	A2 Case AF	A2 Control AF	p-Value
1874633	196	68275196	A/G	0.486	0.551	0.0292
1846060	13311	68288311	G/A	0.416	0.468	0.0792

dbSNP rs#	Position in Figure 4	Chromosome Position	A1/A2 Allele	A2 Case AF	A2 C ntrol AF	p-Value
451352	14486	68289486	C/T	0.474	0.411	0.0365
355468	14691	68289691	A/T	0.839	0.839	0.9913
355469	15551	68290551	C/G	0.089	0.072	0.3028
355470	17702	68292702	T/C	0.077	0.059	0.2261
355471	17872	68292872	T/C	0.476	0.442	0.2613
191650	19588	68294588	T/C	0.122	0.103	0.3282
355472	19910	68294910	T/A	0.491	0.568	0.0114
1874635	20006	68295006	A/C	0.206	0.238	0.2083
1497430	20575	68295575	A/G	0.389	0.476	0.0039
2254659	21092	68296092	G/A	0.554	0.587	0.2664
3822197	22830	68297830	C/T	0.028	0.018	0.2999
2632453	23455	68298455	A/G	0.866	0.895	0.1407
2646282	23716	68298716	G/A	0.137	0.090	0.0146
2646285	23890	68298890	T/G	0.400	0.335	0.0269
768244	24001	68299001	C/T	0.299	0.286	0.6333
724199	24995	68299995	G/A	0.446	0.374	0.0150
1187960	27282	68302282	T/C	0.071	0.060	0.4859
1187961	27779	68302779	C/T	0.499	0.549	0.0968
355518	29099	68304099	C/G	0.432	0.491	0.0473
355519	31185	68306185	A/G	0.095	0.076	0.2836
355511	33994	68308994	C/T	0.450	0.361	0.0030
451397	34942	68309942	T/C	0.442	0.512	0.0210
355513	35137	68310137	C/G	0.385	0.334	0.0748
355514	36538	68311538	T/C	0.423	0.479	0.0596
355515	37139	68312139	C/T	0.422	0.362	0.0395
1056789	37358	68312358	G/A	0.494	0.539	0.1409
2646290	38828	68313828	A/G	0.393	0.337	0.0559
190255	39469	68314469	T/C	0.459	0.514	0.0664
355466	40233	68315233	T/C	0.404	0.468	0.0328
355465	40472	68315472	A/T	0.481	0.547	0.0281
2646292	41679	68316679	C/T	0.422	0.370	0.0820
2632454	41682	68316682	G/A	0.914	0.936	0.1705
1056787	42831	68317831	A/G	0.909	0.860	0.0112
CENPC1_SNP1	42976	68317976	A/G	0.367	0.306	0.0322
173317	44128	68319128	A/G	0.087	0.080	0.6745
451344	44195	68319195	C/T	0.366	0.307	0.0392
355510	46769	68321769	G/A	0.487	0.514	0.3645
355508	47363	68322363	G/C	0.086	0.070	0.3357
451391	48843	68323843	C/T	0.440	0.370	0.0171
355500	52574	68327574	A/G	0.874	0.904	0.1103
355499	52602	68327602	A/G	0.874	0.884	0.5959
355498	53212	68328212	A/G	0.477	0.528	0.0932
1187974	53781	68328781	C/G	0.563	0.540	0.4558
355493	54710	68329710	A/T	0.950	0.932	0.2013
2632456	55808	68330808	G/A	0.091	0.074	0.3234
1825790	57987	68332987	T/A	0.043	0.067	0.0709
355475	58556	68333556	C/A	0.252	0.199	0.0343
1391110	59148	68334148	T/A	0.696	0.679	0.5418
1442557	59286	68334286	G/C	0.458	0.523	0.0306
355478	60217	68335217	A/G	0.314	0.371	0.0474
189579	60412	68335412	G/T	0.008	0.002	0.1543
355480	60753	68335753	C/T	0.905	0.910	0.7624
355481	60791	68335791	T/G	0.974	0.979	0.5823
355483	61524	68336524	A/G	0.371	0.414	0.1461
355485	62543	68337543	T/C	0.487	0.541	0.0732
2646267	62825	68337825	A/G	0.368	0.312	0.0520
2646268	62826	68337826	A/C	0.306	0.239	0.0123
355486	62857	68337857	C/T	0.438	0.375	0.0316

dbSNP rs#	Position in Figure 4	Chromosome Position	A1/A2 Allele	A2 Case AF	A2 Control AF	p-Value
355487	63400	68338400	T/C	0.468	0.559	0.0031
355488	63960	68338960	T/A	0.533	0.454	0.0090
355489	64307	68339307	A/G	0.367	0.324	0.1291
451376	64539	68339539	A/G	0.873	0.871	0.9287
1353626	65728	68340728	A/G	0.356	0.383	0.3657
2632450	66000	68341000	G/A	0.256	0.259	0.9210
2646269	66521	68341521	T/G	0.084	0.062	0.1648
2276945	68185	68343185	C/T	0.459	0.510	0.0866
3775861	69643	68344643	G/A	0.532	0.521	0.7150
1403151	74909	68349909	C/A	0.739	0.801	0.0148
1843833	82973	68357973	T/G	0.920	0.939	0.2355
1843831	83039	68358039	T/C	0.032	0.040	0.5196
3806810	85713	68360713	A/G	0.078	0.058	0.1942
3775862	86873	68361873	T/C	0.744	0.765	0.4224
1962700	90293	68365293	T/G	0.733	0.739	0.8308
2046601	91810	68366810	T/G	0.080	0.073	0.6571
2171386	92609	68367609	A/G	0.685	0.662	0.4056
2046599	92884	68367884	G/A	0.717	0.755	0.1540
355490			A/T	0.495	0.548	0.0763

[0270] Figure 16 shows the proximal SNPs in and around the *ICAM* region for females. The position of each SNP on the chromosome is presented on the x-axis. The y-axis gives the negative logarithm (base 10) of the p-value comparing the estimated allele in the case group to that of the control group. The minor allele frequency of the control group for each SNP designated by an X or other symbol on the graphs in Figure 16 can be determined by consulting Table 22. By proceeding down the Table from top to bottom and across the graphs from left to right the allele frequency associated with each symbol shown can be determined.

[0271] To aid the interpretation, multiple lines have been added to the graph. The broken horizontal lines are drawn at two common significance levels, 0.05 and 0.01. The vertical broken lines are drawn every 20kb to assist in the interpretation of distances between SNPs. Two other lines are drawn to expose linear trends in the association of SNPs to the disease. The light gray line (or generally bottom-most curve) is a nonlinear smoother through the data points on the graph using a local polynomial regression method (W.S. Cleveland, E. Grosse and W.M. Shyu (1992) Local regression models. Chapter 8 of Statistical Models in S eds J.M. Chambers and T.J. Hastie, Wadsworth & Brooks/Cole.). The black line (or generally top-most curve, *e.g.*, see peak in left-most graph just to the left of position 92150000) provides a local test for excess statistical significance to identify regions of association. This was created by use of a 10kb sliding window with 1kb step sizes. Within each window, a chi-square goodness of fit test was applied to compare the proportion of SNPs that were significant at a test wise level of 0.01, to the proportion that would be expected by chance alone (0.05 for the methods used here). Resulting p-values that were less than 10^{-8} were truncated at that value.

[0272] Finally, the gene or genes present in the loci region of the proximal SNPs as annotated by Locus Link (http address: www.ncbi.nlm.nih.gov/LocusLink/) are provided on the graph. The exons and introns of the genes in the covered region are plotted below each graph at the appropriate chromosomal positions. The gene boundary is indicated by the broken horizontal line. The exon positions are shown as thick, unbroken bars. An arrow is place at the 3' end of each gene to show the direction of transcription.

Additional Genotyping

[0273] In addition to the CENCP1 incident SNP, another SNP (rs1056787) was genotyped in the discovery cohort and found to be significantly associated with breast cancer with a p-value of 0.0266. See Table 25.

[0274] The methods used to verify and genotype the proximal SNP of Table 15 are the same methods described in Examples 1 and 2 herein. The PCR primers and extend primers used in these assays are provided in Table 11 and Table 12, respectively.

Table 23

dbSNP rs#	Second PCR primer	First PCR primer
1056787	ACGTTGGATGCATTTCATATTTTGTAGATC	ACGTTGGATGTCTCAGCCCTCTGATAAAAC

Table 24

dbSNP rs#	Extend Primer	Term Mix
1056787	GGATACAAGTTATGCTTTGATAG	ACT

[0275] Table 13, below, shows the case and control allele frequencies along with the p-values for the SNPs genotyped. The disease associated allele of column 4 is in bold and the disease associated amino acid of column 5 is also in bold. The chromosome position provided corresponds to NCBI's Build 33.

Table 25: Genotpying Results

dbSNP rs#	Position in Figure 4	Chromo-some Position	Alleles (A1/A2)	Amino Acid Change	AF F case	AF F control	p-value	Odds Ratio
1056787	42831	68317831	A/G	D389G	A = 0.030 G = 0.970	A = 0.110 G = 0.890	**0.0266**	1.640

Example 7

In Vitro Production of Target Polypeptides

[0276] cDNA is cloned into a pIVEX 2.3-MCS vector (Roche Biochem) using a directional cloning method. A cDNA insert is prepared using PCR with forward and reverse primers having 5' restriction site tags (in frame) and 5-6 additional nucleotides in addition to 3' gene-specific portions, the latter of which is typically about twenty to about twenty-five base pairs in length. A Sal I restriction site is introduced by the forward primer and a Sma I restriction site is introduced by the reverse primer. The ends of PCR products are cut with the corresponding restriction enzymes (*i.e.*, Sal I and Sma I) and the products are gel-purified. The pIVEX 2.3-MCS vector is linearized using the same restriction enzymes, and the fragment with the correct sized fragment is isolated by gel-purification. Purified PCR product is ligated into the linearized pIVEX 2.3-MCS vector and *E. coli* cells transformed for plasmid amplification. The newly constructed expression vector is verified by restriction mapping and used for protein production.

[0277] *E. coli* lysate is reconstituted with 0.25 ml of Reconstitution Buffer, the Reaction Mix is reconstituted with 0.8 ml of Reconstitution Buffer; the Feeding Mix is reconstituted with 10.5 ml of Reconstitution Buffer; and the Energy Mix is reconstituted with 0.6 ml of Reconstitution Buffer. 0.5 ml of the Energy Mix was added to the Feeding Mix to obtain the Feeding Solution. 0.75 ml of Reaction Mix, 50 μl of Energy Mix, and 10 μg of the template DNA is added to the *E. coli* lysate.

[0278] Using the reaction device (Roche Biochem), 1 ml of the Reaction Solution is loaded into the reaction compartment. The reaction device is turned upside-down and 10 ml of the Feeding Solution is loaded into the feeding compartment. All lids are closed and the reaction device is loaded into the RTS500 instrument. The instrument is run at 30°C for 24 hours with a stir bar speed of 150 rpm. The pIVEX 2.3 MCS vector includes a nucleotide sequence that encodes six consecutive histidine amino acids on the C-terminal end of the target polypeptide for the purpose of protein purification. Target polypeptide is purified by contacting the contents of reaction device with resin modified with Ni^{2+} ions. Target polypeptide is eluted from the resin with a solution containing free Ni^{2+} ions.

Example 8

Cellular Production of Target Polypeptides

[0279] Nucleic acids are cloned into DNA plasmids having phage recombination cites and target polypeptides are expressed therefrom in a variety of host cells. Alpha phage genomic DNA contains short sequences known as attP sites, and *E. coli* genomic DNA contains unique, short sequences known as attB sites. These regions share homology, allowing for integration of phage DNA into *E. coli* via directional, site-specific recombination using the phage protein Int and the *E. coli* protein IHF. Integration produces two new att sites, L and R, which flank the inserted prophage DNA. Phage excision from *E. coli* genomic DNA can also be accomplished using these two proteins with the addition of a second phage protein, Xis. DNA vectors have been produced where the integration/excision process is modified to allow for the directional integration or excision of a target DNA fragment into a backbone vector in a rapid *in vitro* reaction (Gateway™ Technology (Invitrogen, Inc.)).

[0280] A first step is to transfer the nucleic acid insert into a shuttle vector that contains attL sites surrounding the negative selection gene, ccdB (*e.g.* pENTER vector, Invitrogen, Inc.). This transfer process is accomplished by digesting the nucleic acid from a DNA vector used for sequencing, and to ligate it into the multicloning site of the shuttle vector, which will place it between the two attL sites while removing the negative selection gene ccdB. A second method is to amplify the nucleic acid by the polymerase chain reaction (PCR) with primers containing attB sites. The amplified fragment then is integrated into the shuttle vector using Int and IHF. A third method is to utilize a topoisomerase-mediated process, in which the nucleic acid is amplified via PCR using gene-specific primers with the 5' upstream primer containing an additional CACC sequence (*e.g.*, TOPO® expression kit (Invitrogen, Inc.)). In conjunction with Topoisomerase I, the PCR amplified fragment can be cloned into the shuttle vector via the attL sites in the correct orientation.

[0281] Once the nucleic acid is transferred into the shuttle vector, it can be cloned into an expression vector having attR sites. Several vectors containing attR sites for expression of target polypeptide as a native polypeptide, N-fusion polypeptide, and C-fusion polypeptides are commercially available (*e.g.*, pDEST (Invitrogen, Inc.)), and any vector can be converted into an expression vector for receiving a nucleic acid from the shuttle vector by introducing an insert having an attR site flanked by an antibiotic resistant gene for selection using the standard methods described above. Transfer of the nucleic acid from the shuttle vector is accomplished by directional recombination using Int, IHF, and Xis (LR clonase). Then the desired sequence can be transferred to an expression vector by carrying out a one hour incubation at room temperature with Int, IHF, and Xis, a ten minute incubation at 37°C with proteinase K, transforming bacteria and allowing expression for one hour, and then plating on selective

media. Generally, 90% cloning efficiency is achieved by this method. Examples of expression vectors are pDEST 14 bacterial expression vector with att7 promoter, pDEST 15 bacterial expression vector with a T7 promoter and a N-terminal GST tag, pDEST 17 bacterial vector with a T7 promoter and a N-terminal polyhistidine affinity tag, and pDEST 12.2 mammalian expression vector with a CMV promoter and neo resistance gene. These expression vectors or others like them are transformed or transfected into cells for expression of the target polypeptide or polypeptide variants. These expression vectors are often transfected, for example, into murine-transformed a adipocyte cell line 3T3-L1, (ATCC), human embryonic kidney cell line 293, and rat cardiomyocyte cell line H9C2.

Example 9

Haplotype analysis of the *KIAA0783* locus

[0282] Markers rs1681290, rs220097, rs3801435, and rs2883140 are significantly associated with breast cancer at the allele and genotype levels ($P < 0.05$). Strong LD is observed between markers 1681290, 220097, 3801435, and 2883140 ($r^2 > 0.90$). Pearson chi-squared statistics indicate that haplotypes are significantly associated with breast cancer. Haplotypes TTGCGG, CTGCGG, and TCATAT contribute most to the aggregate test statistic. Odds ratios and score tests indicate that individuals with the TTGCGG and CTGCGG haplotypes are significantly less likely to have breast cancer, while individuals with the TCATAT haplotype are slightly more likely to be affected than individuals with other haplotypes.

Statistics

[0283] Chi-squared statistics are estimated to assess whether 1) alleles and genotypes are associated with breast cancer status and 2) marker genotype frequencies deviate significantly from Hardy-Weinberg equilibrium (HWE). Haplotype frequencies and relative frequencies are estimated, as well as several statistics (r^2, D', and p-value) that gauge the extent and stability of linkage disequilibrium between markers in each region. Chi-squared statistics and score tests are estimated to determine whether reconstructed haplotypes are significantly associated with breast cancer status ($P < 0.05$). P-values are estimated for 1) the full set of reconstructed haplotypes and 2) a reduced set that excludes haplotypes with observed frequencies less than 10. Results are presented by chromosome order.

Results

Summary Statistics: Alleles and Genotypes

SNP Locations

SNP.ID	Type	Location
218981	Proximal	10720511
1681284	Proximal	10739011
1681290	Proximal	10741656
220097	Incident	10759860
3801435	Proximal	10771563
2883140	Proximal	10806368

Allele by GYNGroup

	N	Case (N=510)	Control (N=538)	Test Statistic
218981:T	1028	47%(232)	45%(239)	Chi-square=0.68 d.f.=1 P=0.41
1681284:C	1032	56%(276)	50%(267)	Chi-square=3.51 d.f.=1 P=0.0608
1681290:A	1018	72%(352)	63%(330)	Chi-square=8.92 d.f.=1=0.00282
220097:C	996	29%(139)	38%(196)	Chi-square=8.03 d.f.=1P=0.00461|
3801435:G	1018	28%(138)	38%(200)	Chi-square=9.69 d.f.=1P=0.00185|
2883140:T	1012	73%(351)	62%(330)	Chi-square=12.78 d.f.=1 P<0.001|

Genotype by GYNGroup

	N	Case (N=255)	Control (N=269)	Test Statistic
218981:CC	514	27%(67)	27%(73)	Chi-square=2.41 d.f.=2 P=0.299
CT		51%(126)	56%(151)	
TT		22%(53)	16%(44)	
1681284:TT	516	19%(48)	26%(70)	Chi-square=3.77 d.f.=2 P=0.152
TC		50%(124)	48%(129)	
CC		31%(76)	26%(69)	
1681290:GG	509	9%(21)	16%(41)	Chi-square=8.64 d.f.=2 P=0.0133
GA		40%(98)	43%(114)	
AA		52%(127)	41%(108)	

	N	Case (N=255)	Control (N=269)	Test Statistic
220097:TT	498	50%(119)	40%(104)	Chi-square=8.06 d.f.=2 P=0.0177
TC		42%(99)	45%(116)	
CC		8%(20)	15%(40)	
3801435:AA	509	51%(124)	40%(107)	Chi-square=9.78 d.f.=2 P=0.0075
AG		41%(100)	44%(118)	
GG		8%(19)	15%(41)	
2883140:GG	506	8%(19)	16%(42)	Chi-square=12.14 d.f.=2 P=0.00231
GT		39%(93)	44%(116)	
TT		54%(129)	40%(107)	

Genotype QC: Test of Hardy-Weinberg Proportions

All

	A.freq	D	ChiSq	Pvalue
218981	0.543	-0.01990	3.290	0.0697
1681284	0.526	0.00564	0.263	0.6080
1681290	0.670	0.01170	1.430	0.2320
220097	0.664	0.00584	.351	.5530
3801435	0.667	0.00585	.355	.5510
2883140	0.675	.01360	.970	.1610

Control

	A.freq	D	ChiSq	Pvalue
218981	0.554	-0.03380	5.010	0.0252
1681284	0.502	0.01030	0.453	0.5010
1681290	0.627	0.01470	1.050	0.3050
220097	0.620	0.00904	0.393	0.5310
3801435	0.624	0.01190	0.684	0.4080
2883140	0.625	0.01700	1.410	0.2350

Summary Statistics: Linkage Disequilibrium

PHASE Haplotype Frequencies

	H.freq	H.relfreq
CCATAT	91	0.089
CCGCGG	4	0.004
CTACGG	5	0.005
CTACGT	1	0.001
CTATAT	142	0.138
CTGCAG	1	0.001
CTGCAT	2	0.002
CTGCGG	300	0.292
CTGCGT	10	0.010
CTGTAT	1	0.001
TCACGG	1	0.001
TCATAG	1	0.001
TCATAT	443	0.432
TTATAT	3	0.003
TTGCGG	21	0.020

Linkage Disequilibrium Between Markers

r^2

x	218981	1681284	1681290	220097	3801435	2883140
218981	1.000	0.603	0.311	0.316	0.311	0.292
1681284	0.603	1.000	0.524	0.532	0.525	0.498
1681290	0.311	0.524	1.000	0.965	0.952	0.914
220097	0.316	0.532	0.965	1.000	0.987	0.940
3801435	0.311	0.525	0.952	0.987	1.000	0.944
2883140	0.292	0.498	0.914	0.940	0.944	1.000

D'

	218981	1681284	1681290	220097	3801435	2883140
218981	1.000	0.803	0.728	0.725	0.724	0.715
1681284	0.803	1.000	0.978	0.972	0.972	0.966
1681290	0.728	0.978	1.000	0.996	0.982	0.969
220097	0.725	0.972	0.996	1.000	1.000	0.995
3801435	0.724	0.972	0.982	1.000	1.000	0.991
2883140	0.715	0.966	0.969	0.995	0.991	1.000

P-value

	218981	1681284	1681290	220097	3801435	2883140
218981	1	0	0	0	0	0
1681284	0	1	0	0	0	0
1681290	0	0	1	0	0	0
220097	0	0	0	1	0	0
3801435	0	0	0	0	1	0
2883140	0	0	0	0	0	1

Haplotype by GYNGroup

All Haplotypes

	Case	Case(%)	Case.X^2	Control	Control(%)	Control.X^2	OR	ln.OR
CTGCAG	0	0.00	0.48	1	0.10	0.44	0.0000	-Inf
TCACGG	0	0.00	0.48	1	0.10	0.44 0.0000	-Inf	
TCATAG	0	0.00	0.48	1	0.10	0.44	0.0000	-Inf
TTATAT	0	0.00	1.44	3	0.29	1.33	0.0000	-Inf
TTGCGG	1	0.10	8.17	20	1.95	7.53	0.0491	-3.0139
CCGCGG	1	0.10	0.44	3	0.29	0.40	0.3327	-1.1005

	Case	Case(%)	Case.X^2	Control	Control(%)	Control.X^2	OR	ln.OR
CTACGG	2	0.19	0.07	3	0.29	0.06	0.6660	-0.4065
CTGCGG	129	12.57	1.53	171	16.67	1.41	0.7191	-0.3298
CCATAT	43	4.19	0.01	48	4.68	0.01	0.8913	-0.1151
CTGCAT	1	0.10	0.00	1	0.10	0.00	1.0000	0.0000
TCATAT	230	22.42	1.45	213	20.76	1.34	1.1029	0.0979
CTATAT	76	7.41	0.92	66	6.43	0.85	1.1636	0.1515
CTGCGT	7	0.68	1.01	3	0.29	0.93	2.3425	0.8512
CTACGT	1	0.10	0.56	0	0.00	0.52	Inf	Inf
CTGTAT	1	0.10	0.56	0	0.00	0.52	Inf	Inf

Pearson Chi-squared Test = 33.8392, DF = 14, P-value = 0.002177

Permutation Test P-value = 0.01

PHASE Haplotypes (Low Frequency Excluded)

	Case	Case(%)	Case.X^2	Control	Control(%)	Control.X^2	OR	ln.OR
TTGCGG	1	0.10	8.23	20	1.99	7.68	0.0491	-3.0139
CTGCGG	129	12.81	1.72	171	16.98	1.61	0.7183	-0.3309
CCATAT	43	4.27	0.02	48	4.77	0.02	0.8912	-0.1152
TCATAT	230	22.84	1.23	213	21.15	1.14	1.1034	0.0984
CTATAT	76	7.55	0.81	66	6.55	0.76	1.1639	0.1518
CTGCGT	7	0.70	0.98	3	0.30	0.91	2.3427	0.8513

Pearson Chi-squared Test = 25.1157, DF = 5, P-value = 0.0001323

haplo.score Haplotypes

	Hap.Freq	Score	P. X^2	P.Sim
TTGCGG	0.0203	-3.7664	0.0002	0.0001
TTATAT	0.0063	-2.5040	0.0123	0.0097
CTGCGG	0.2947	-2.0103	0.0444	0.0438
CCATAT	0.0902	-0.3982	0.6905	0.7174
CTATAT	0.1318	1.4254	0.1540	0.1538
CTGCGT	0.0084	1.5778	0.1146	0.1243
TCATAT	0.4342	2.3889	0.0169	0.0180

Global Score = 27.2432, DF = 7, Global P.X^2 = 3e-04, Global P.Sim = 1e-04

Example 10

Haplotype analysis of the *CENPC1* locus

[0284] Each SNP noted below is significantly associated with breast cancer at allele level ($P < 0.05$). rs355510 maintains a significant relationship with disease at the genotype level. Near-complete LD is observed across the entire region. Pearson chi-squared statistics demonstrate that haplotypes CCAC and TTGT are significantly associated with breast cancer after low frequency haplotypes are removed from the analysis. Odds ratios and score tests indicate that individuals with the CCAC haplotype are significantly less likely to have breast cancer, while individuals with the TTGT haplotype are at moderately increased risk for disease vs. individuals with other haplotypes.

Statistics

[0285] Chi-squared statistics are estimated to assess whether 1) alleles and genotypes are associated with breast cancer status and 2) marker genotype frequencies deviate significantly from Hardy-Weinberg equilibrium (HWE). Haplotype frequencies and relative frequencies are estimated, as well as several statistics (r^2, D', and p-value) that gauge the extent and stability of linkage disequilibrium between markers in each region. Chi-squared statistics and score tests are estimated to determine whether reconstructed haplotypes are significantly associated with breast cancer status ($P < 0.05$). P-values are estimated for 1) the full set of reconstructed haplotypes and 2) a reduced set that excludes haplotypes with observed frequencies less than 10. Results are presented by chromosome order.

Results

Summary Statistics: Alleles and Genotypes

SNP Locations

SNP.ID	Type	Location
GP04.071927035	Proximal	68289486
355511	Proximal	68308994
355510	Incident	68321769
355487	Proximal	68338400

Allele by GYNGroup

	N	Case (N=508)	Control (N=536)	Test Statistic
GP04.071927035:T	1022	46% (225)	39% (207)	Chi-square=5.13 d.f.=1 P=0.0235
355511:C	1010	55% (264)	61% (321)	Chi-square=4.34 d.f.=1 P=0.0371
355510:A	1004	54% (261)	62% (321)	Chi-square=6.27 d.f.=1 P=0.0123
355487:C	992	54% (262)	61% (311)	Chi-square=5.1 d.f.=1 P=0.0239

Genotype by GYNGroup

	N	Case (N=254)	Control (N=268)	Test Statistic
GP04.071927035:CC	511	28%(69)	37%(98)	Chi-square=5.33 d.f.=2 P=0.0695
CT		52%(127)	48%(129)	
TT		20%(49)	15%(39)	
355511:TT	505	20%(48)	14%(38)	Chi-square=4.47 d.f.=2 P=0.107
TC		51%(124)	49%(129)	
CC		29%(70)	37%(96)	
355510:GG	502	20%(49)	15%(38)	Chi-square=6.52 d.f.=2 P=0.0383
GA		52%(125)	47%(123)	
AA		28%(68)	38%(99)	
355487:TT	496	20%(48)	15%(37)	Chi-square=5.35 d.f.=2 P=0.069
TC		52%(126)	48%(123)	
CC		28%(68)	37%(94)	

Genotype QC: Test of Hardy-Weinberg Proportions

All

	A.freq	D	ChiSq	Pvalue
GP04.071927035	0.577	-0.00599	0.303	0.582
355511	0.579	-0.00630	0.337	0.562
355510	0.577	-0.00599	0.303	0.582
355487	0.577	-0.00599	0.303	0.582

Control

	A.freq	D	ChiSq	Pvalue
GP04.071927035	0.609	-0.00420	0.0814	0.775

355511	0.611-0.00653	0.1970	0.657	
355510	0.609	-0.00420	0.0814	0.775
355487	0.611	-0.00271	0.0340	0.854

Summary Statistics: Linkage Disequilibrium

PHASE Haplotype Frequencies

	H.freq	H.relfreq
CCAC	581	0.576
CCAT	1	0.001
TCGT	2	0.002
TTGC	1	0.001
TTGT	423	0.420

Linkage Disequilibrium Between Markers

r^2

	GP04.071927035	355511	355510	355487
GP04.071927035	1.000	0.992	1.000	0.992
355511	0.992	1.000	0.992	0.984
355510	1.000	0.992	1.000	0.992
355487	0.992	0.984	0.992	1.000

D'

	GP04.071927035	355511	355510	355487
GP04.071927035	1.000	1.000	1.000	0.996
355511	1.000	1.000	1.000	0.996
355510	1.000	1.000	1.000	0.996

355487	0.996	0.996	0.996	1.000

P-value

	GP04.071927035	355511	355510	355487
GP04.071927035	1	0	0	0
355511	0	1	0	0
355510	0	0	1	0
355487	0	0	0	1

Haplotype by GYNGroup

PHASE Haplotypes (All)

	Case	Case(%)	Case.X^2	Control	Control(%)	Control.X^2	OR	ln.OR
TTGC	0	0.00	0.48	1	0.10	0.44	0.0000	-Inf
CCAC	262	25.99	1.03	319	31.65	0.95	0.7586	-0.2763
TCGT	1	0.10	0.00	1	0.10	0.00	1.0000	0.0000
TTGT	220	21.83	1.41	203	20.14	1.30	1.1071	0.1017
CCAT	1	0.10	0.56	0	0.00	0.52	Inf	Inf

Pearson Chi-squared Test = 6.6985, DF = 4, P-value = 0.1527

Permutation Test P-value = 0.56

PHASE Haplotypes (Low Frequency Excluded)

	Case	Case(%)	Case.X^2	Control	Control(%)	Control.X^2	OR	ln.OR
CCAC	262	26.10	1.03	319	31.77	0.95	0.7582	-0.2768
TTGT	220	21.91	1.41	203	20.22	1.30	1.1072	0.1018

Pearson Chi-squared Test = 4.4162, DF = 1, P-value = 0.0356

haplo.score Haplotypes

	Hap.Freq	Score	P.X^2	P.Sim

CCAC	0.5772	-2.3513	0.0187	0.0168
TTGT	0.4208	2.2111	0.0270	0.0249

Global Score = 7.5085, DF = 2, Global P.X^2 = 0.0234, Global P.Sim = 0.0117

[0286] Citation of the above publications or documents is not intended as an admission that any of the foregoing is pertinent prior art, nor does it constitute any admission as to the contents or date of these publications or documents. U.S. patents and other publications referenced herein are hereby incorporated by reference.

[0287] Modifications may be made to the foregoing without departing from the basic aspects of the invention. Although the invention has been described in substantial detail with reference to one or more specific embodiments, those of skill in the art will recognize that changes may be made to the embodiments specifically disclosed in this application, yet these modifications and improvements are within the scope and spirit of the invention, as set forth in the claims which follow. All publications or patent documents cited in this specification are incorporated herein by reference as if each such publication or document was specifically and individually indicated to be incorporated herein by reference.